Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

На правах рукописи

Осокин Антон Александрович

Субмодулярная релаксация в задаче минимизации энергии марковского случайного поля

Специальность 01.01.09 — дискретная математика и математическая кибернетика

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
к.ф.-м.н. Д. П. Ветров

Москва — 2014

Содержание

Введение

В рамках данной диссертационной работы разработан новый подход к решению задачи поиска наиболее вероятных конфигураций марковских случайных полей: субмодулярная релаксация (submodular relaxation, SMR). Проведено теоретическое и экспериментальное исследование предложенного подхода, а также сравнение его с аналогами.

Актуальность темы. В связи с бурным развитием цифровых технологий в последние 10-15 лет появилась необходимость в решении большого количества задач, связанных с обработкой высокоуровневой информации: изображений, видео, звука, и т. д. Важной особенностью таких данных является наличие большого числа внутренних зависимостей или структуры. Например, на фотореалистичных изображениях цвета соседних точек (пикселей) чаще всего сильно коррелированы, на видеопоследовательностях коррелированы цвета не только соседних пикселей, но и цвета одного и того же пикселя на соседних кадрах. При анализе таких данных многие классические методы машинного обучения и распознавания образов, ориентированные на работу с выборками независимых случайных величин, оказываются неприменимыми или не показывают хороших результатов.

Большинство задач распознавания заключается в предсказании неизвестных величин на основе наблюдаемых. Можно выделить два типа задач распознавания, связанных со структурированными данными:

1. предсказания всех ненаблюдаемых величин осуществляются независимо;
2. предсказания ненаблюдаемых величин осуществляются согласованно.

Задачи первого типа обладают структурой на уровне описания данных, но не обладают ей на уровне выхода алгоритма распознавания. Примерами задач первого типа являются задачи классификации изображений (для изображения определить к какому классу оно относится: внутри помещения или снаружи, название страны, в которой сделана фотография), задачи определения говорящего по звуковой последовательности (в предположении, что всё время говорит один человек), и др. Задачи второго типа обладают структурой как на уровне описания данных, так и на уровне выхода алгоритма распознавания. Примерами таких задач являются сегментация

изображений (сопоставление метки класса каждому пикселю), отслеживание (трекинг) объекта на видео, восстановление произнесённой фразы по звукозаписи, и т. д.

В решении задач первого типа в настоящее время доминируют нейронные сети нового поколения [83, 48] (глубинное обучение, deep learning). Данная парадигма делает попытку получить по данным признаковое описание, содержащее представление внутренних закономерностей. Методы, основанные на глубинном обучении, показывают лучшие на сегодняшний день результаты при решении большого числа прикладных задач (например, для задачи классификации изображений по самой большой открытой базе изображений ImageNet [75]).

Для решения задач второго типа большой популярностью пользуется аппарат, так называемых, графических моделей [20, 129]. Данный подход делает попытку напрямую моделировать закономерности данных, затрагивающие как признаковое описание, так и результат распознавания. Обычно под графической моделью понимается вероятностная модель, задающая совместное распределение большого количества переменных, структура зависимостей в котором задаётся при помощи графа или гиперграфа. Важным отличием методов, относящихся к графическим моделям, от методов для решения задач типа 1 является то, что сложной является не только задача обучения модели (настройка параметров по наблюдаемым данным), но и задача распознавания нового объекта по уже обученной модели.

Один из наиболее популярных подходов к математической формулировке и решению задачи распознавания второго типа основан на поиске моды совместного апостериорного распределения неизвестных переменных (maximum a posteriori estimation, MAP-inference). Задача поиска моды является задачей оптимизации (либо непрерывной, либо дискретной). Часто распределение представляет собой произведение большого числа множителей – факторов, и работать с ним в таком виде неудобно. В этом случае берут отрицательный логарифм апостериорного распределения и переходят к эквивалентной задаче минимизации. Минимизируемую функцию обычно называют энергией[1]. Несмотря на то что задача минимизации энергии в общем случае является NP-трудной [113], на практике её приближённые решения получать существенно проще, чем, например, приближённо вычислять апостериорные маргинальные распределения.

Задачи минимизации энергии часто возникают в качестве подзадачи при решении задачи настройки параметров модели по наблюдаемым данным. Наиболее известным методом, в котором возникает подзадача минимизации энергии, является структурный метод опорных векторов (structured support vector machine, SSVM) [122, 124]. SSVM часто используется на практике для решения задач второго типа, т. к. в ряде случаев превосходит альтернативные методы как по качеству, так и по скорости работы [95, 100].

[1] Термин энергия используется из-за связи с понятием потенциальной энергии из статистической физики [92].

В рамках данной работы изучается задача минимизации энергии, в которой, во-первых, все переменные энергии дискретны (задача минимизации энергии является задачей дискретной оптимизации), и, во-вторых, существует компактное представление энергии в виде суммы слагаемых, каждое из которых зависит от небольшого числа переменных (опр. 1.1). Задачам минимизации таких энергии уделяется внимание как отечественными [8, 2, 3, 4, 9], так и зарубежными учёными (например, в работах [21, 54]).

Задача минимизации энергии дискретных переменных появилась достаточно давно: в отечественной литературе она исследовалась ещё в 70-х в работах М. И. Шлезингера [8]. В западной литературе первые работы (из известных автору) появились в начале 80-х: Гиман и Гиман [41], Блейк и Циссерман [21] сформулировали задачу именно в том виде, в котором она часто рассматривается сейчас, а также предложили алгоритм имитации отжига (simulated annealing) для её решения. Вехой в развитии данной задачи стали работы Перла [99], в которых были сформулированы алгоритмы передачи сообщений и оформилось понимание того, что если граф энергии не содержит циклов, то задача может быть решена точно за полиномиальное время.

Важный класс функций, допускающих минимизацию за полиномиальное время, – субмодулярные функции бинарных переменных – был известен среди специалистов по дискретной оптимизации ещё в 60-х годах [46]. В конце 80-х годов Грейг [44] и др. впервые использовали подход, основанный на минимизации энергии при помощи построения минимальных разрезов графов, в задаче подавления шума на бинарных изображениях. В начале 00-х годов работы Бойкова и др. [24, 26], Колмогорова и Заби [68] положили начало активному использованию разрезов графов в компьютерном зрении. В качестве примеров задач, решаемых при помощи минимизации энергии, можно привести стерео-сопоставление [24], сегментацию изображений [26, 114], восстановление неизвестных областей изображения (inpainting) [94], сегментацию трёхмерных объектов [85, 12]. По мере роста числа приложений подхода происходил рост и размеров задач, и сложности используемых потенциалов. Например, для задачи сегментации изображений было разработано большое число потенциалов, позволяющих учитывать сложные высокоуровневые свойства объектов [61, 86, 93, 88].

Наиболее изучена задача минимизации в случае парно-сепарабельных энергий, т. е. энергий, являющихся суммами слагаемых, зависящих не более чем от двух переменных. Экспериментальные исследования [120, 54] показывают, что для случая парно-сепарабельных энергий существует большое число алгоритмов, позволяющих решать многие практические задачи с требуемой точностью. В случае же не парно-сепарабельных энергий (энергий с потенциалами, зависящими от более 2-х переменных) арсенал существующих методов гораздо более скромен. Существующие методы либо позволяют минимизировать потенциалы очень специальных ви-

дов [61, 32, 80], либо работают недостаточно быстро [104, 71], либо обладают одновременно обоими недостатками [86, 93].

Целью данной диссертационной работы является разработка метода решения задачи минимизации энергии с потенциалами высоких порядков, который должен быть, во-первых, применим к энергиям достаточно общего вида, а, во-вторых, должен превосходить существующие аналоги на задачах минимизации энергии некоторых типов.

Методы исследования. Для достижения поставленной цели был выбран подход, основанный на релаксации Лагранжа ограничений, затрудняющих решение задачи. Частным случаем этого подхода является двойственная декомпозиция (dual decomposition), применённая для задач минимизации энергии в работах Комодакиса и др.[73, 71], Вудфорда и др. [135], Зонтага и др. [117]. В настоящей диссертации используется вариант релаксации Лагранжа, выходящий за рамки двойственной декомпозиции. Также для разработки метода активно используются алгоритмы построения разрезов графов [23] и их динамических расширений [59].

Основные положения, выносимые на защиту:

1. Новый подход для решения задачи минимизации энергии: субмодулярная релаксация.
2. Доказательства эквивалентности разработанного подхода ряду существующих аналогов в случаях парно-сепарабельных энергий и энергий с потенциалами высоких порядков специального вида.
3. Алгоритм покоординатного подъема для максимизации нижней оценки, построенной в рамках подхода субмодулярной релаксации, применимый в случае ассоциативных парно-сепарабельных энергий.
4. Экспериментальное исследование всех разработанных методов, содержащее их сравнение с существующими аналогами.

Научная новизна настоящей диссертации заключается в разработке нового подхода к решению задачи минимизации энергии; получении ряда теоретических результатов, включающих в себя формулировки эквивалентных задач линейного программирования; экспериментальном исследовании предложенного подхода, состоящем в сравнении с аналогами и демонстрации применимости на практике.

Теоретическая значимость настоящей работы состоит в том, что закрыт целый ряд вопросов, возникающих при появлении нового семейства нижних оценок (субмодулярная релаксация), основанных на релаксации Лагранжа. В частности, приведен точный вид задачи линейного

программирования, решение которой эквивалентно наилучшей нижней оценке; проведен теоретический анализ свойств семейства оценок; разработаны алгоритмы поиска наилучшей нижней оценки в рамках предложенного семейства.

Практическая значимость настоящей работы состоит в том, что разработанный алгоритм на ряде важных прикладных задач (энергии с разреженными потенциалами высоких порядков) оказывается быстрее аналогов, и, соответственно, является шагом в направлении более широкого использования алгоритмов минимизации энергии на практике.

Степень достоверности. Достоверность результатов обеспечивается доказательствами теорем и подробными описаниями экспериментов, допускающими воспроизводимость.

Апробация работы. Результаты настоящей работы неоднократно докладывались на семинаре группы байесовских методов машинного обучения кафедры математических методов прогнозирования, ВМК МГУ, а также докладывались на следующих конференциях:

1. Международная конференция по компьютерному зрению и распознаванию образов (Computer Vision and Pattern Recognition, CVPR), 2010. [31]

2. Международная конференция «Интеллектуализация обработки информации» (ИОИ), 2010. [128, 77]

3. Международная конференция по компьютерному зрению и распознаванию образов (Computer Vision and Pattern Recognition, CVPR), 2011. [97]

4. Всероссийская конференция «Математические методы распознавания образов» (ММРО), 2011. [7]

5. Международная конференция «Системы обработки нейроинформации» (Neural Information Processing Systems, NIPS), секция «Дискретная оптимизация в машинном обучении» (discrete optimization in machine learning, DISCML), 2011. [127]

6. Международная конференция «Системы обработки нейроинформации» (Neural Information Processing Systems, NIPS), 2012. [33]

7. Европейская конференция по компьютерному зрению (European Conference on Computer Vision, ECCV), секция «Модели высоких порядков и глобальные ограничения в компьютерном зрении» (Higher-Order Models and Global Constraints in Computer Vision), 2012. [96]

8. Международная конференция по компьютерному зрению и распознаванию образов (Computer Vision and Pattern Recognition, CVPR), 2013. [64]

Публикации. Основные результаты по теме диссертации изложены в 11 печатных изданиях [7, 31, 32, 33, 64, 76, 77, 96, 97, 127, 128], 7 из которых входят в список изданий, рекомендованных ВАК [31, 32, 33, 64, 76, 96, 97], 4 – сборники докладов конференций [7, 77, 127, 128]. Отдельные результаты настоящей работы включались в отчёты по проектам РФФИ 08-01-00405, 12-01-31254, 12-01-00938, 12-01-33085, и по проекту МК 3827.2010.9.

Личный вклад диссертанта заключается в выполнении основного объёма теоретических и экспериментальных исследований, изложенных в диссертационной работе, включая разработку теоретических моделей, методик экспериментальных исследований, проведение исследований, анализ и оформление результатов в виде публикаций и научных докладов. К личному вкладу диссертанта не относится формулировка и доказательство теоремы 2.

Объём и структура работы. Диссертация состоит из оглавления, введения, пяти глав, заключения, списка иллюстраций (19 п.), списка таблиц (2 п.), списка литературы (139 п.) и одного приложения. Общий объём работы составляет – 121 стр.

Краткое содержание работы по главам.

В главе 1 вводятся используемые обозначения, приводится формальная постановка задачи. Далее приводится краткое описание существующих методов решения задачи минимизации энергии. В заключении данной главы содержится подробное описание нескольких методов оптимизации энергии, имеющих наиболее близкое отношение к настоящей работе:

1. алгоритмы передачи сообщений для точного решения задачи минимизации энергии в случае ациклического графа [99] и фактор-графа [20];
2. алгоритмы точной минимизации парно-сепарабельных энергий, зависящих от бинарных [68] и многозначных [30] переменных, основанные на построении минимальных разрезов графов;
3. приближённые алгоритмы минимизации энергии, основанные на итеративном построении разрезов графов [24];
4. приближённые алгоритмы минимизации энергии, основанные на линейной релаксации дискретной задачи и двойственной декомпозиции [73, 71].

В главе 2 приведено формальное описание предлагаемого подхода субмодулярной релаксации. Сначала изложен частный случай подхода – субмодулярная декомпозиция [97], применимый в случае парно-сепарабельных ассоциативных энергий. Далее излагается подход в общем виде. Завершает данную главу описание расширения подхода на случай несубмодулярного лагранжиана, а также описание способа учёта глобальных линейных ограничений на индикаторные переменные.

В главе 3 проводится теоретическое исследование предлагаемого подхода в общем виде и двух его частных случаев: парно-сепарабельные ассоциативные энергии и парно-сепарабельные неассоциативные энергии. Во всех случаях формулируется задача линейного программирования, решению которой эквивалентен каждый метод (теоремы 1-5). Аналогичные результаты приводятся и для случаев наличия линейных глобальных ограничений (теоремы 6-8).

Глава 4 содержит теоретическое исследование, посвящённое различным вопросам, связанным с максимизацией нижней оценки, возникающей при субмодулярной релаксации. Проводится теоретическое исследование свойств точки максимума, формулируются понятия сильной и слабой согласованности. Приводится анализ применимости различных методов выпуклой оптимизации конкретной функции, возникающей в рамках предлагаемого подхода. Формулируется метод покоординатного подъёма для максимизации нижней оценки. Доказывается сходимость метода и ряд свойств точки сходимости (теорема 9).

Глава 5 посвящена экспериментальному исследованию подхода субмодулярной релаксации. Проводится сравнение различных методов оптимизации, а также сравнение с аналогами для случаев парно-сепарабельных ассоциативных функций, разреженных потенциалов высокого порядка, неассоциативных парно-сепарабельных потенциалов. Показывается применимость предлагаемого подхода для задач сегментации изображений и сегментации магнитограмм Солнца.

Благодарности. Автор выражает благодарность своему научному руководителю Дмитрию Петровичу Ветрову за внимание, активное участие в работе и воспитание; соруководителю Дмитрию Александровичу Кропотову за понимание и личный пример; жене, родителям и брату за поддержку и терпение; соавторам Юрию Бойкову, Эндрю Делонгу, Владимиру Колмогорову, Пушмиту Коли за советы и сотрудничество; а также студенту Александру Новикову за плодотворные дискуссии.

1. Задача минимизации дискретных энергий

В этой главе вводится нотация, используемая в рамках данной работы, приводится формальная постановка решаемой задачи (задачи минимизации энергии). Проводится краткий обзор существующих методов решения этой задачи, а также подробный обзор группы методов, имеющих наиболее близкое отношение к настоящей работе: алгоритмы передачи сообщений на графах и фактор-графах, алгоритмы минимизации энергии, основанные на построении разрезов графов, релаксационные алгоритмы.

1.1. Нотация и постановка задачи

В этом разделе даются базовые определения и вводится нотация, используемая в данной работе. За основу взята нотация, использованная в работах [133, 17].

Рассмотрим гиперграф $\mathcal{G} = (\mathcal{V}, \mathcal{C})$, где $\mathcal{V}$ – конечное множество вершин, $\mathcal{C}$ – конечное мультимножество гиперрёбер – подмножеств множества вершин $\mathcal{V}$: $\mathcal{C} \subseteq 2^{\mathcal{V}}$.

Пусть каждой вершине гиперграфа $i \in \mathcal{V}$ соответствует переменная x_i, принимающая значения из конечного непустого множества меток $\mathcal{P}$. Для любого гиперребра $C \in \mathcal{C}$ символом $\boldsymbol{x}_C$ обозначим кортеж переменных, индексы которых принадлежат множеству C: $\boldsymbol{x}_C = (x_i \mid i \in C)$[1], а символом $\mathcal{X}_C$ – совместное множество значений этих переменных: $\mathcal{X}_C = \underset{i \in C}{\times} \mathcal{P}$, $\mathcal{X}_i = \mathcal{P}$. При обозначении кортежа всех переменных из $\mathcal{V}$ и множества значений этих переменных индекс $\mathcal{V}$ будем опускать: $\boldsymbol{x}$, $\mathcal{X}$.

Определение 1. *Назовём энергией, заданной на гиперграфе $\mathcal{G}$, функционал следующего вида:*

$$E(\boldsymbol{x}) = \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{C \in \mathcal{C}} \theta_C(\boldsymbol{x}_C). \tag{1.1}$$

[1] Здесь и далее будем считать, что на множестве вершин $\mathcal{V}$ задан полный порядок (нумерация $1, \ldots, |\mathcal{V}|$). При переходе от подмножества C множества $\mathcal{V}$ (неупорядоченного набора вершин) к кортежу $\boldsymbol{x}_C$ (упорядоченному набору переменных) упорядочивание элементов будем проводить в соответствии с этим порядком.

Здесь функционалы $\theta_i : \mathcal{P} \to \mathbb{R}$ *называются унарными потенциалами, а функционалы* $\theta_C : \mathcal{X}_C \to \mathbb{R}$ *– потенциалами, заданными на гиперрёбрах.*

Для каждой переменной $i \in \mathcal{V}$ множество $\mathcal{C}(i) \subseteq \mathcal{C}$ содержит гиперрёбра, инцидентные переменой i: $\mathcal{C}(i) = \{C \mid C \in \mathcal{C},\ i \in C\}$.

Порядком потенциала θ_C называется количество вершин, входящее в множество C. Порядок унарных потенциалов равен 1. Потенциалы порядка 2 назовём *парными*, потенциалы больших порядков – *потенциалами высоких порядков*. Потенциалы, для которых мощности гиперребер равны общему количеству переменных $|\mathcal{V}|$, назовём *глобальными*. Для обозначения значений унарных потенциалов будем использовать символы $\theta_{ix_i} = \theta_i(x_i) = \theta_{\{i\}}(\boldsymbol{x}_{\{i\}})$, парных – $\theta_{ij,x_ix_j} = \theta_{ij}(x_i, x_j) = \theta_{\{i,j\}}(\boldsymbol{x}_{\{i,j\}})$, потенциалов произвольных порядков – $\theta_{C,\boldsymbol{x}_C} = \theta_C(\boldsymbol{x}_C)$.

Энергии, состоящие только из унарных и парных потенциалов, назовём *парно-сепарабельными*. Парно-сепарабельные энергии будем записывать следующим способом:

$$E(\boldsymbol{x}) = \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{\{i,j\} \in \mathcal{E}} \theta_{ij}(x_i, x_j). \tag{1.2}$$

Здесь $\mathcal{E}$ – множество *рёбер* (гиперрёбер мощности 2). В этой нотации множество гиперрёбер $\mathcal{C}$ состоит из гиперрёбер порядка 2: $\mathcal{C} = \{\{i,j\} \mid \{i,j\} \in \mathcal{E}\}$. Множество вершин $\mathcal{V}$ и множество рёбер $\mathcal{E}$ образуют граф $(\mathcal{V}, \mathcal{E})$, поэтому часто говорят, что парно-сепарабельная энергия задается графом.

Данная работа посвящена задаче минимизации энергии $E(\boldsymbol{x})$ (1.1) по дискретным переменным $\boldsymbol{x}$:

$$\min_{\boldsymbol{x} \in \mathcal{X}} E(\boldsymbol{x}). \tag{1.3}$$

Задача (1.3) является задачей дискретной оптимизации. Известно, что для произвольного гиперграфа $\mathcal{G}$ и произвольных потенциалов задача (1.3) является NP-трудной [113, 24]. Данная работа посвящена разработке приближённых методов решения данной задачи. Раздел 1.3.1 содержит описание частных случаев, когда задача (1.3) может быть решена точно, а раздел 1.3.2 посвящён приближённым методам решения задачи (1.3), имеющим наиболее близкое отношение к данной работе.

В дальнейшем изложении, для удобства, будем использовать индикаторную нотацию. Для каждой метки $p \in \mathcal{P}$ и переменной $x_i, i \in \mathcal{V}$ введем индикатор – бинарную переменную $y_{ip} := [x_i = p]$[2]. Энергия (1.1) в индикаторной нотации выглядит следующим образом:

$$E_I(\boldsymbol{y}) = \sum_{i \in \mathcal{V}} \sum_{p \in \mathcal{P}} \theta_{ip} y_{ip} + \sum_{C \in \mathcal{C}} \sum_{\boldsymbol{d} \in \mathcal{X}_C} \theta_{C,\boldsymbol{d}} \prod_{i \in C} y_{id_i}. \tag{1.4}$$

[2]Здесь и далее символы $[A]$, где A – логическое выражение, соответствуют скобке Айверсона: $[A] := 1$, если A истинно; $[A] := 0$, если A ложно.

Здесь символы d_i, где $i \in \mathcal{V}$, обозначают ту метку из кортежа меток $\boldsymbol{d}$, которая соответствует переменной x_i из исходного множества переменных (существует и единственна, когда $i \in C$).

Очевидно, что задача безусловной оптимизации (1.3) эквивалентна задаче условной оптимизации, записанной в терминах индикаторных переменных:

$$\min_{\boldsymbol{y}} \quad E_I(\boldsymbol{y}) \tag{1.5}$$

$$\text{s.t.} \quad \boldsymbol{y} \in \{0,1\}^{|\mathcal{P}||\mathcal{V}|}, \tag{1.6}$$

$$\sum_{p \in \mathcal{P}} y_{ip} = 1, \quad \forall i \in \mathcal{V}. \tag{1.7}$$

Ограничения (1.6), (1.7) гарантируют, что каждой переменной x_i исходного набора переменных, ставится в соответствие вектор из $|\mathcal{P}| - 1$ нуля и ровно одной единицы, по которому исходную метку из множества $\mathcal{P}$ можно однозначно восстановить. Ограничения (1.6) будем называть *целочисленностью*, а ограничения (1.7) – *целостностью*.

В данной работе задачи условной оптимизации вида (1.5)-(1.7) будет удобно записывать следующим образом: $\min_{\boldsymbol{y} \in (1.6),(1.7)} E_I(\boldsymbol{y})$.

1.2. Энергии и марковские случайные поля

Энергия, заданная на гиперграфе $\mathcal{G}$ (1.1), имеет непосредственное отношение к марковским случайным полям (Markov random fields, MRF) и часто называется энергией MRF.

Действительно, рассмотрим распределение Гиббса над множеством $\mathcal{X}$, где вероятность конфигурации $\boldsymbol{x}$ определяется следующим образов:

$$\mathsf{P}(\boldsymbol{x}) = \frac{1}{Z} \exp\left(-\frac{1}{T} E(\boldsymbol{x})\right).$$

Здесь $T > 0$ – параметр распределения, называемый *температурой*. Z – нормировочная константа, равная $\sum_{\boldsymbol{x} \in \mathcal{X}} \exp\left(-\frac{1}{T} E(\boldsymbol{x})\right)$.

Говорят, что данное распределение задаёт Марковское случайное поле (вероятностную графическую модель), которая факторизуется согласно гиперграфу $\mathcal{G}$:

$$\mathsf{P}(\boldsymbol{x}) = \frac{1}{Z} \prod_{i \in \mathcal{V}} \psi_i(x_i) \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}_C).$$

Здесь ψ_i и ψ_C – факторы MRF, равные $\exp(-\theta_i(x_i)/T)$ и $\exp(-\theta_C(\boldsymbol{x}_C)/T)$, соответственно.

Задача минимизации энергии (1.1) эквивалентна задаче поиска наиболее вероятного состояния MRF, а при наличии наблюдаемых переменных (Conditional Random Field, CRF) [81] – задаче поиска максимума апостериорного распределения (maximum a posteriori probability estimate, MAP).

1.3. Методы минимизации энергии

В этом разделе приводится обзор существующих методов решения задачи минимизации энергии (1.3).

В общем виде задача является NP-трудной [113, 24], а значит решить её в общем случае за полиномиальное время не представляется возможным[3]. Тем не менее, существует ряд частных случаев, когда задача полиномиально разрешима:

- граф (фактор-граф, см. опр. 2, для энергии с потенциалами высоких порядков) не содержит циклов [99, 20];
- энергия является парно-сепарабельной и субмодулярной [44, 68, 30];
- энергия представляет собой парно-сепарабельную модель Изинга, а граф планарен [109];
- граф энергии имеет небольшую ширину [82, 129].

В разделе 1.3.1 приведены описания первых двух случаев, как наиболее известных и имеющих наибольшее отношение к настоящей работе.

Для ситуаций, когда задача минимизации энергии является NP-трудной, существует целый ряд приближённых методов. Можно выделить несколько основных групп приближённых методов:

- алгоритмы, делающие шаги (метод покоординатного спуска (ICM) [19], алгоритмы, основанные на разрезах графов [24, 87]);
- релаксационные алгоритмы (методы покоординатного подъёма [8, 5, 132, 42, 117], алгоритмы, основанные на декомпозиции [130, 65, 73]);
- алгоритмы передачи сообщений (LBP [138], TRW [130]);
- стохастические методы (имитация отжига [41], методы сэмплирования [20]);
- комбинаторные алгоритмы (метод ветвей и границ [118, 55]).

В разделе 1.3.2 приведены описания нескольких методов, имеющих наиболее близкое отношение к теме настоящей работы.

Существует несколько работ по сравнению методов различных видов на широком спектре энергий, возникающих в прикладных задачах [120, 54].

Также существует большое количество методов, разработанных для потенциалов специальных видов (часто специфичных конкретным прикладным задачам): потенциалы, обеспечивающие ограничения на глобальные статистики разметок [135, 76, 32, 80, 91], связность [93] и

[3] Если только P не равно NP.

априорные распределения на форму объекта [39, 125, 86, 89] в задаче сегментации изображений, и др.

1.3.1. Частные случаи, допускающие точные решения

Как уже было сказано выше, задача минимизации энергии (1.1) в общем виде является NP-трудной. Тем не менее, существует несколько важных частных случаев, когда задачу можно решить точно за полиномиальное время. Выделим два способа ограничить количество степеней свободы задачи: ограничения на структуру гиперграфа при произвольных потенциалах, ограничения на вид потенциалов при произвольной структуре гиперграфа.

1.3.1.1. Ограничения на структуру гиперграфа

1.3.1.1.1. Парно-сепарабельная энергия с графом без циклов. Рассмотрим задачу минимизации парно-сепарабельной энергии (1.2), заданной на графе $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Пусть граф $\mathcal{G}$ не содержит циклов и состоит из одной компоненты связности (является деревом).

Назовём сообщением от вершины i к вершине j следующий вектор:

$$\mu_{i\to j}(x_j) = \min_{x_i} \quad \theta_{ij}(x_i, x_j) + \theta_i(x_i) + \sum_{\substack{\{k,i\}\in\mathcal{E}\\ k\neq j}} \mu_{k\to i}(x_i) \tag{1.8}$$

Сообщение от вершины i к вершине j рекуррентно зависит от сообщений из всех соседей вершины i, кроме j, в вершину i.

Алгоритм передачи сообщений выбирает некоторое начальное приближение значений сообщений, после чего пересчитывает их в некотором порядке, называемом *расписанием*, до сходимости. Покажем, при каком расписании алгоритм передачи сообщений точно решает задачу минимизации энергии (1.2), заданной на ациклическом графе.

Без ограничения общности предположим, что граф связный. Выделим произвольную вершину графа ℓ и назовём её корнем. Пронумеруем все вершины графа $\mathcal{G}$, кроме ℓ, в таком порядке, что для дерева обхода графа $\mathcal{G}$ в глубину, начиная с вершины ℓ, все потомки стоят раньше родителей.

Будем передавать сообщения от вершины к её родителю (согласно дереву обхода в глубину) в соответствии с введённой нумерацией. При этом либо в выражении (1.8) в сумме не будет элементов (если текущая вершина является листом), либо все слагаемые суммы уже будут вычислены. Всего нужно вычислить $|\mathcal{V}| - 1$ сообщение. Данное расписание проиллюстрировано на рис. 1.1.

При таком расписании переданное сообщение от вершины i к вершине j соответствует минимальной энергии поддерева, «висящего» на ребре $\{i, j\}$ со стороны i, если переменная x_j принимает фиксированное значение. Минимум всей энергии, при этом, может быть вычислен при помощи суммирования всех сообщений, входящих в корень ℓ, и унарного потенциала корня:

$$\min_{\boldsymbol{x}\in\mathcal{X}} E(\boldsymbol{x}) = \min_{x_\ell\in\mathcal{P}} \theta_\ell(x_\ell) + \sum_{k:\{k,\ell\}\in\mathcal{E}} \mu_{k\to\ell}(x_\ell). \tag{1.9}$$

Конфигурация, на которой достигается минимум энергии, может быть найдена при помощи рекуррентного взятия аргминимума в формулах (1.9) и (1.8).

Описанный выше алгоритм является прямым обобщением алгоритма динамического программирования для случая, если граф $\mathcal{G}$ имеет вид цепочки. Примером моделей-цепочек являются скрытые марковские модели (hidden Markov models, HMM). При работе с HMM алгоритм динамического программирования обычно называют алгоритмом Витерби [20].

Отметим, что сложность алгоритма передачи сообщений при вычислении по формулам (1.8) напрямую составляет $O(|\mathcal{V}||\mathcal{P}|^2)$, что может быть достаточно медленно при большом количестве меток (например, в задаче обнаружении человека на изображении количество меток равно количеству возможных позиций каждой части тела [37]). Тем не менее, оказывается, что для ряда типов парных потенциалов выражение (1.8) можно вычислять за линейное по числу меток время. Например, если парные потенциалы являются потенциалами Поттса ($\theta_{ij}(x_i, x_j) = c_{ij}[x_i \neq x_j]$, $c_{ij} \in \mathbb{R}$), то сообщения можно вычислить следующим образом:

$$\mu_{i\to j}(x_j) = \min\left(\hat{\theta}_{i\to j}(x_j),\ \min_{x_i\in\mathcal{P}\setminus\{x_j\}} \hat{\theta}_{i\to j}(x_i) + c_{ij}\right),$$

где $\hat{\theta}_{i\to j}(x_i) = \theta_i(x_i) + \sum_{\substack{\{k,i\}\in\mathcal{E}\\ k\neq j}} \mu_{k\to i}(x_i)$. Такие вычисления требуют уже $O(|\mathcal{V}||\mathcal{P}|)$ операций. В работах [35, 36] предложен способ быстрых вычислений сообщений потенциалов более общего вида, основанных на быстром вычислении преобразований расстояний (distance transforms).

1.3.1.1.2. Обобщения на случай потенциалов высоких порядков. Алгоритм передачи сообщений несложно обобщить на случай энергий с потенциалами высоких порядков. Для этого будем использовать понятие фактор-графа.

Определение 2. Фактор-графом, *соответствующим энергии* (1.1), *заданной на гиперграфе* $\mathcal{G} = (\mathcal{V}, \mathcal{C})$, *назовём граф* $\hat{\mathcal{G}} = (\hat{\mathcal{V}}, \hat{\mathcal{E}})$, *в котором вершины* $\hat{\mathcal{V}}$ *соответствуют как вершинам, так и факторам (слагаемым в выражении для энергии) исходного гиперграфа* $\mathcal{G}$, *а рёбра соединяют вершины фактора* C *и переменной* x_i, *только когда фактор* C *зависит от переменной* i: $i \in C$.

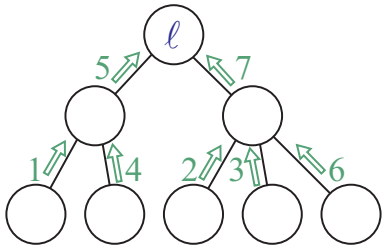


Рисунок 1.1.: Расписание алгоритма передачи сообщений, позволяющее точно решить задачу минимизации парно-сепарабельной энергии, заданной на ацикличном графе из 8 вершин. Вершина ℓ выделена и является корнем дерева обхода в глубину. Стрелки вдоль рёбер показывают направления сообщений, задействованных при вычислении минимума энергии с данным корнем. Цифры возле стрелок показывают одну из возможных последовательностей передач сообщений, позволяющую точно решить задачу.

Граф, определённый таким образом, является двудольным (рёбра соединяют только вершины-факторы с вершинами-переменными). Вершины фактор-графа, соответствующие переменным, будем индексировать при помощи индексов исходных переменных $i \in \mathcal{V}$; вершины фактор-графа, соответствующие переменным, – при помощи гиперрёбер $C \in \mathcal{C}$.

Стоит отметить, что по энергии, записанной формулой (1.1), фактор-граф строится неоднозначно, поскольку слагаемые можно разными способами объединять в факторы, и, более того, можно объединять несколько переменных в одну. Рис. 1.2 показывает примеры фактор-графов, построенных для одной энергии разными способами. Здесь и далее будем считать, что все преобразования с энергией проведены до формирования гиперграфа $\mathcal{G}$, а фактор-граф $\hat{\mathcal{G}}$ построен «естественным образом», описанным выше.

Определим сообщения для фактор-графов. Выделим сообщения двух видов: от факторов к вершинам (1.10) и от вершин к факторам (1.11).

$$\mu_{C\to i}(x_i) = \min_{\boldsymbol{x}_{C\setminus\{i\}}} \theta_C(\boldsymbol{x}_C) + \sum_{j\in C\setminus\{i\}} \mu_{j\to C}(x_j), \tag{1.10}$$

$$\mu_{i\to C}(x_i) = \theta_i(x_i) + \sum_{\substack{A:A\in\mathcal{C},\\ i\in A,\ A\neq C}} \mu_{A\to i}(x_i). \tag{1.11}$$

Аналогично случаю парно-сепарабельного графа, алгоритм передачи сообщений, начиная с некоторой инициализации, пересчитывает сообщения по формулам (1.10) и (1.11) в соответствии с некоторым порядком - расписанием.

Если фактор-граф не содержит циклов, то легко показать, что существует расписание, позволяющее точно решить задачу минимизации энергии. Такое расписание строится аналогично расписанию для случая парно-сепарабельных графов, а именно как передача сообщений от листьев к корню согласно дереву поиска в глубину, построенному от некоторой выделенной вершины ℓ. Пример одного из таких расписаний приведен на рис. 1.3. После того, как все необходимые

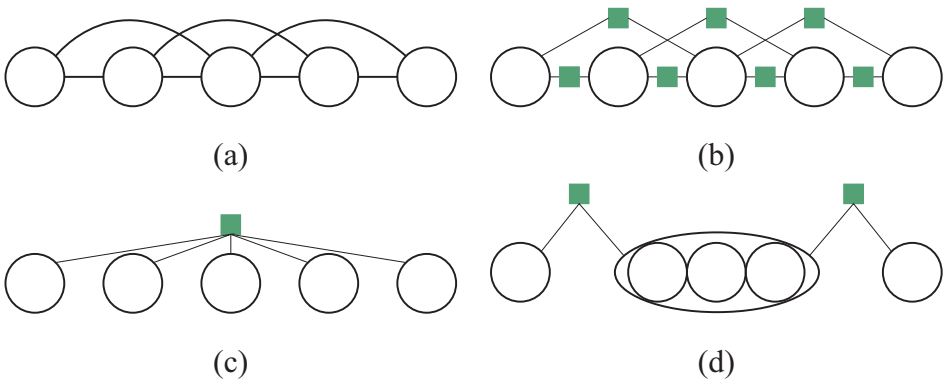


Рисунок 1.2.: Фактор-графы, построенные для энергии, заданной на графе (a). (b) соответствует фактор графу, построенному «естественным способом». Данный фактор-граф содержит циклы. (c) соответствует фактор-графу той же энергии, но построенному при помощи другой группировки слагаемых по факторам. Данный фактор-граф не содержит циклов, но, при такой группировке, фактически, произошёл отказ от факторизации модели. (d) содержит фактор-граф, построенный при помощи объединения нескольких переменных в одну. Данный фактор-граф одновременно не содержит циклов и содержит информацию (хотя и не всю) о факторизации.

сообщения вычислены минимум энергии может быть вычислен аналогично (1.9):

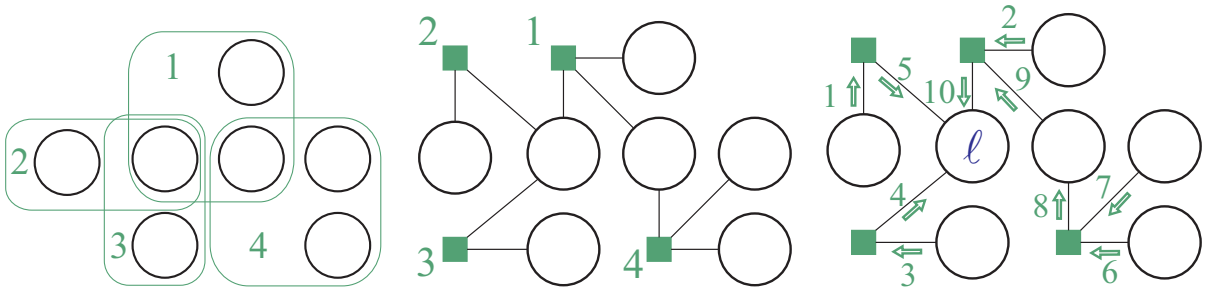
$$\min_{\boldsymbol{x}\in\mathcal{X}} E(\boldsymbol{x}) = \min_{x_\ell\in\mathcal{P}} \theta_\ell(x_\ell) + \sum_{\substack{C:\, C\in\mathcal{C},\\ \ell\in C}} \mu_{C\to\ell}(x_\ell). \tag{1.12}$$

Нетрудно видеть, что сложность алгоритма передачи сообщений линейна по количеству вершин и факторов, но экспоненциальна (формула (1.10)) по размеру факторов. Аналогично случаю парно-сепарабельных энергий для факторов некоторых специальных видов сообщения можно вычислять существенно быстрее. Примерами таких ситуаций являются энергии с глобальными потенциалами специальных видов [121], а также алгоритм кластеризации «распространение близости» (affinity propagation) [40].

1.3.1.2. Ограничения на вид потенциалов

1.3.1.2.1. Парно-сепарабельная субмодулярная энергия с бинарными переменными.

Рассмотрим задачу минимизации парно-сепарабельной энергии (1.2), заданной на произвольном графе $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Пусть все переменные бинарны: $\mathcal{P} = \{0, 1\}$. Функции, отображающие булев куб на множество действительных чисел, часто называют *псевдо-булевыми*. Известно, что если не вводить никаких дополнительных ограничений, то задача минимизации парно-сепарабельной псевдо-булевой функции является NP-трудной [68]. Однако, если все парные потенциалы такой функции удовлетворяют условию *субмодулярности*, то задачу можно решить за полиномиальное время при помощи сведения к задачам построения максимального потока и минимального



(a) (b) (c)

Рисунок 1.3.: Обобщение алгоритма передачи сообщений на случай энергии с потенциалами высоких порядков на примере энергии из 7 вершин и 4-х факторов. (a) – структура факторов энергии; области показывают вершины, объединённые факторами. (b) – ацикличный фактор-граф, построенный для данной энергии. (c) – расписание алгоритма передачи сообщений; стрелки вдоль рёбер показывают направления сообщений, задействованных при вычислении минимума энергии при выделенном корне ℓ. Цифры возле стрелок показывают одну из возможных последовательностей передач сообщений, позволяющую точно решить задачу.

разреза в графе. Данный результат был известен ещё в середине 60-х годов [46, 22], но был переоткрыт в начале нулевых В. Колмогоровым и др. [68]. С тех пор алгоритмы минимизации энергии, основанные на использовании алгоритмов построения разрезов графов (graph cuts) активно используются в задачах компьютерного зрения и машинного обучения.

Определение субмодулярности. Рассмотрим одно из центральных понятий теории псевдобулевых функций – *субмодулярность*.

Определение 3. *Функция бинарных переменных $E(\boldsymbol{x})$ называется субмодулярной, если для любых двух разметок $\boldsymbol{x}^1$ и $\boldsymbol{x}^2$ выполнено условие*

$$E(\boldsymbol{x}^1 \vee \boldsymbol{x}^2) + E(\boldsymbol{x}^1 \wedge \boldsymbol{x}^2) \leqslant E(\boldsymbol{x}^1) + E(\boldsymbol{x}^2), \tag{1.13}$$

где операции «$\vee$» и «$\wedge$» – поэлементная дизъюнкция (максимум) и конъюнкция (минимум), соответственно[4].

[4] Часто определение субмодулярности формулируется для функций, определённых на всех подмножествах множества переменных $\mathcal{V}$: $g : 2^{\mathcal{V}} \to \mathbb{R}$. Условие (1.13) в этом случае записывается следующим образом:

$$g(A \cup B) + g(A \cap B) \leqslant g(A) + g(B), \tag{1.14}$$

где $A, B \in 2^{\mathcal{V}}$. Если каждому подмножеству $\mathcal{V}$ взаимнооднозначно поставить в соответствие вектор индикаторных переменных длины $|\mathcal{V}|$, то условия (1.14) и (1.13) определяют эквивалентные классы функций.

Утверждение 1. *Парно-сепарабельная функция бинарных переменных $E(\boldsymbol{x})$ является субмодулярной, тогда и только тогда, когда для каждого парного потенциала выполнено следующее условие:*

$$\theta_{ij}(0,0) + \theta_{ij}(1,1) \leqslant \theta_{ij}(0,1) + \theta_{ij}(1,0). \tag{1.15}$$

Доказательство. Достаточность следует из того, что в случае функции двух переменных условия (1.15) и (1.13) совпадают, а также того, что сумма субмодулярных является субмодулярной. Для доказательства необходимости можно рассмотреть в качестве $\boldsymbol{x}^1$ и $\boldsymbol{x}^2$ вектора, совпадающие во всех компонентах, кроме i и j, и содержащие значения 0 и 1 на i-й и j-й позициях в разном порядке. □

Часто, если речь идет о парно-сепарабельных функциях, именно условие (1.15) используют в качестве определения субмодулярности.

Соответствие минимизации энергии разрезу графа. Рассмотрим парно-сепарабельную энергию бинарных переменных, в которой потенциалы удовлетворяют следующим ограничениям:

- унарные потенциалы неотрицательны: $\forall i \in \mathcal{V} \quad \theta_i(0) \geqslant 0,\ \theta_i(1) \geqslant 0$;
- парные потенциалы неотрицательны и равны 0 при равенстве связанных переменных: $\forall \{i,j\} \in \mathcal{E} \quad \theta_{ij}(0,0) = \theta_{ij}(1,1) = 0,\ \theta_{ij}(0,1) \geqslant 0,\ \theta_{ij}(1,0) \geqslant 0$.

В этом случае энергию можно записать в следующем виде:

$$E(\boldsymbol{x}) = \sum_{i \in \mathcal{V}} (x_i \theta_i(1) + (1 - x_i)\theta_i(0)) + \sum_{\{i,j\} \in \mathcal{E}} (x_i(1 - x_j)\theta_{ij}(1,0) + x_j(1 - x_i)\theta_{ij}(0,1)) + \theta_0. \tag{1.16}$$

По энергии (1.16) построим ориентированный граф $\bar{\mathcal{G}} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$, st-разрез которого будет соответствовать присвоению значений переменным $\boldsymbol{x}$[5]:

- Множество вершин $\bar{\mathcal{V}} = \mathcal{V} \cup \{s,t\}$, где s и t – две дополнительные вершины: исток и сток, соответственно.
- Множество дуг $\bar{\mathcal{E}}$ строится следующим образом: каждому ребру $\{i,j\} \in \mathcal{E}$ ставим в соответствие две дуги (i,j) и (j,i) (*нетерминальные дуги*), каждой вершине $i \in \mathcal{V}$ ставим в соответствие дуги (s,i) и (i,t) (*терминальные дуги*).
- st-разрезом графа будем считать разбиение всех вершин множества $\bar{\mathcal{V}}$ на два непересекающихся множества $\mathcal{S}$ и $\mathcal{T}$, $\mathcal{S} \cap \mathcal{T} = \emptyset$, $\mathcal{S} \cup \mathcal{T} = \bar{\mathcal{V}}$, такое что $s \in \mathcal{S}$ и $t \in \mathcal{T}$. Будем считать, что множество истока $\mathcal{S}$ соответствует значению переменных 0, а множество стока $\mathcal{T}$ – значению 1.

[5] Основные определения и факты, связанные с разрезами графов и потоками в сетях, даны в дополнении А.

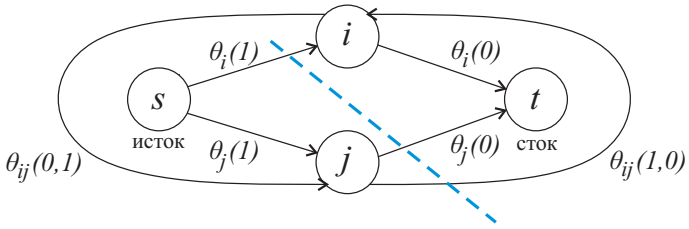


Рисунок 1.4.: Граф, построенный для минимизации энергии от двух переменных x_i, x_j. Разрез, отображенный пунктирной линией, соответствует присваиванию $x_i = 1$, $x_j = 0$. Величина разреза составляет $\theta_i(1) + \theta_j(0) + \theta_{ij}(1,0)$.

- Веса терминальных дуг: $c(s,i) = \theta_i(1)$, $c(i,t) = \theta_i(0)$, где $i \in \mathcal{V}$.
- Веса нетерминальных дуг: $c(i,j) = \theta_{ij}(0,1)$, $c(j,i) = \theta_{ij}(1,0)$, где $\{i,j\} \in \mathcal{E}$, и, по договоренности, $i < j$.
- Присваивание значений переменных по построенному st-разрезу строится следующим образом: $i \in \mathcal{S} \Rightarrow x_i = 0$, $i \in \mathcal{T} \Rightarrow x_i = 1$. Величина разреза совпадает со значением энергии (1.16) при таких значениях переменных $\boldsymbol{x}$. Таким образом, задача минимизации энергии вида (1.16) эквивалента задаче поиска минимального st-разреза в графе $\bar{\mathcal{G}}$.

Пример графа, построенного для энергии, зависящей от 2-х переменных, и его разреза приведен на рис. 1.4.

Репараметризация. Рассмотрим, какие ещё энергии (кроме (1.16)) можно минимизировать при помощи разрезов графов.

На энергию $E(\boldsymbol{x})$ можно смотреть как на функцию. У одной и той же функции может быть несколько различных записей вида (1.2). Назовем преобразование записи энергии $E(\boldsymbol{x})$, не меняющее энергию, как функцию, *репараметризацией*. Рассмотрим несколько видов репараметризаций:

- Вычитание константы из унарного потенциала: $\theta_i(0) := \theta_i(0) - \delta$, $\theta_i(1) := \theta_i(1) - \delta$, $\theta_0 := \theta_0 + \delta$. Здесь $i \in \mathcal{V}$, $\delta \in \mathbb{R}$.
- Изменение парных потенциалов: $\theta_{ij}(p,0) := \theta_{ij}(p,0) - \delta$, $\theta_{ij}(p,1) := \theta_{ij}(p,1) - \delta$, $\theta_i(p) := \theta_i(p) + \delta$. Аналогично $\theta_{ij}(0,p) := \theta_{ij}(0,p) - \delta$, $\theta_{ij}(1,p) := \theta_{ij}(1,p) - \delta$, $\theta_j(p) := \theta_j(p) + \delta$. Здесь $\{i,j\} \in \mathcal{E}$, $\delta \in \mathbb{R}$, $p \in \{0,1\}$.

Легко показать, что описанные преобразования не меняют энергию $E(\boldsymbol{x})$ как функцию.

Рассмотрим, какие парно-сепарабельные энергии общего вида (1.2) при помощи описанных преобразований-репараметризаций можно свести к виду (1.16). Заметим, что применяя репараметризацию унарных потенциалов с $\delta = \min(\theta_i(0), \theta_i(1))$ унарные потенциалы всегда можно

сделать неотрицательными. Таким образом, можно снять все ограничения на унарные потенциалы энергии (1.16).

Рассмотрим, что можно сделать при помощи репараметризаций парных потенциалов. Пусть потенциал ребра $\{i,j\} \in \mathcal{E}$ принимает следующие значения: $\theta_{ij}(0,0) = a$, $\theta_{ij}(1,1) = b$, $\theta_{ij}(0,1) = c$, $\theta_{ij}(1,0) = d$. Применим следующую последовательность преобразований:

1. $\theta_{ij}(0,0) := \theta_{ij}(0,0) - a$, $\theta_{ij}(0,1) := \theta_{ij}(0,1) - a$, $\theta_i(0) := \theta_i(0) + a$;
2. $\theta_{ij}(0,1) := \theta_{ij}(0,1) - c + a$, $\theta_{ij}(1,1) := \theta_{ij}(1,1) - c + a$, $\theta_j(1) := \theta_j(1) + c - a$;
3. $\theta_{ij}(1,1) := \theta_{ij}(1,1) - b + c - a$, $\theta_{ij}(1,0) := \theta_{ij}(1,0) - b + c - a$, $\theta_i(1) := \theta_j(1) + b - c + a$.

Можно убедиться, что данные преобразования приводят к следующим значениям парного потенциала θ_{ij}: $\theta_{ij}(0,0) = \theta_{ij}(1,1) = \theta_{ij}(0,1) = 0$, $\theta_{ij}(1,0) = d+c-a-b$. Если после этого при помощи описанных ранее действий сделать все унарные потенциалы неотрицательными, то энергия будет иметь вид (1.16), тогда и только тогда, когда выполнено условие $d+c-a-b \geqslant 0$. Заметим, что это условие в точности соответствует условию субмодулярности парно-сепарабельной энергии бинарных переменных (1.15). Данное условие вызвано тем, что для полиномиальной разрешимости задач о максимальном потоке и минимальном разрезе пропускные способности дуг графа должны быть неотрицательны.

1.3.1.2.2. Обобщение на случай $|\mathcal{P}|$-значных переменных. Рассмотрим задачу минимизации парно-сепарабельной энергии (1.2), заданной на произвольном графе $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Пусть все переменные принимают одно из K значений: $\mathcal{P} = \{0, \ldots, K-1\}$. Если на множестве меток определён полный порядок и все парные потенциалы являются субмодулярными (обобщение понятия на K-значный случай), то задачу можно решить за полиномиальное время при помощи алгоритмов построения минимального разреза графа. Первой работой в данном направлении была работа Ишикавы [50]; Дарбон расширил применимость метода и сформулировал его в текущем виде [30].

Определение понятия субмодулярности для небинарных переменных аналогично бинарному случаю (опр. 3):

Определение 4. *Функция K-значных переменных $E(\boldsymbol{x})$ называется субмодулярной, если для любых двух разметок $\boldsymbol{x}^1$ и $\boldsymbol{x}^2$ выполнено условие*

$$E(\boldsymbol{x}^1 \vee \boldsymbol{x}^2) + E(\boldsymbol{x}^1 \wedge \boldsymbol{x}^2) \leqslant E(\boldsymbol{x}^1) + E(\boldsymbol{x}^2), \tag{1.17}$$

где операции «$\vee$» и «$\wedge$» – поэлементный максимум и минимум, соответственно. Подразумевается, что эти две операции определены для любых значений переменных, что соответствует введению на множестве полного порядка.

Критерием субмодулярности парно-сепарабельных энергий является субмодулярность всех парных потенциалов (аналогично бинарному случаю, утв. 1).

Введём индикаторные переменные $[x]_{ip}$, $i \in \mathcal{V}$, $p \in \mathcal{P}$, соответствующие линиям уровня переменной x_i:

$$[x]_{ip} = \begin{cases} 0, & x_i \leqslant p, \\ 1, & x_i > p. \end{cases} \tag{1.18}$$

Унарные и парные потенциалы исходной энергии можно выразить через введённые переменные [30]. Унарные потенциалы выражаются следующим образом:

$$\theta_i(x_i) = \sum_{p=0}^{K-1} D_i(p)[x]_{ip} + \theta_{i,0}, \tag{1.19}$$

где $D_i(p) = \theta_{i,p+1} - \theta_{i,p}$, $p \in \{0, \dots K-2\}$. Парные потенциалы можно выразить так:

$$\theta_{ij}(x_i, x_j) = \sum_{p=0}^{K-2}\sum_{q=0}^{K-2} R_{ij}(p,q)[x]_{ip}[x]_{jq} + \sum_{p=0}^{K-2}(D^1_{ij}(p)[x]_{ip} + D^2_{ij}(q)[x]_{jq}) + D^0. \tag{1.20}$$

Здесь

$$\begin{aligned} R_{ij}(p,q) &= \theta_{ij,p+1\,q+1} - \theta_{ij,p+1\,q} - \theta_{ij,p\,q+1} + \theta_{ij,p\,q}, \\ D^1_{ij}(p) &= \theta_{ij,p+1\,0} - \theta_{ij,p\,0}, \\ D^2_{ij}(q) &= \theta_{ij,0\,q+1} - \theta_{ij,0\,q}, \\ D^0 &= \theta_{ij,0\,0}. \end{aligned}$$

Выражения (1.19) и (1.20) позволяют переписать энергию (1.2) через парно-сепарабельную функцию бинарных переменных $[x]_{ip}$. Условие того, что переменные $[x]_{ip}$ являются линиями уровня какой-либо разметки переменной x_i (в смысле определения (1.18)), можно записать как парные потенциалы энергии:

$$M(1 - [x]_{ip})[x]_{i\,p+1}. \tag{1.21}$$

Здесь M – достаточно большое положительное число [30].

Заметим, что все парные потенциалы, появляющиеся в выражениях (1.19), (1.20), (1.21) являются субмодулярными[6], а значит, согласно критерию субмодулярности парно-сепарабельной функции (утв. 1) полученное выражение является субмодулярным. Таким образом, в данном случае можно применить метод минимизации субмодулярных парно-сепарабельных псевдо-булевых функций, описанный в разделе 1.3.1.2.1.

[6]Потенциалы (1.21) субмодулярны, поскольку $M > 0$. Потенциалы (1.20) субмодулярны, поскольку $R_{ij}(p,q) \leqslant 0$, что следует из опр. 4

1.3.1.2.3. Потенциалы высоких порядков. Рассмотрим задачу минимизации энергии с потенциалами высоких порядков (1.1), где все переменные являются бинарными. Известно, что субмодулярные (опр. 3) энергии высоких порядков могут быть точно минимизированы за полиномиальное время. Тем не менее, все известные алгоритмы обладают высокой сложностью, что делает их малоприменимыми на практике (например, алгоритм [51] обладает сложностью $O(|\mathcal{V}|^5)$). Для потенциалов небольших порядков более близкими к практичным оказываются алгоритмы, обладающие не строго полиномиальной сложностью. Например, алгоритм [14], обладающий сложностью $O(|\mathcal{V}|^3 2^M)$, где $M = \max_{C \in \mathcal{C}} |C|$.

Из-за высокой сложности общих алгоритмов на практике часто используются специализированные алгоритмы, применимые лишь к потенциалам определённых видов. Наиболее популярным способом минимизации энергий вида (1.1) является сведение их к парно-сепарабельным энергиям, а именно добавление в энергию новых переменных $\boldsymbol{z}$, так что

$$\min E(\boldsymbol{x}) = \min_{\boldsymbol{x},\boldsymbol{z}} E^*(\boldsymbol{x}, \boldsymbol{z}),$$

где E^* – парно-сепарабельная субмодулярная функция.

Наиболее популярным (и хронологически первым [38]) примером такой редукции является тождество

$$-\prod_{i \in C} x_i = \min_{z \in \{0,1\}} \left(z(|C| - 1) - \sum_{i \in C} z x_i \right).$$

На каждое слагаемое энергии, при котором стоит неположительный коэффициент, вводится дополнительная переменная z, так чтобы относительно переменных $\boldsymbol{x}$ и $\boldsymbol{z}$ функция была парно-сепарабельной и субмодулярной. При этом, минимум энергии E^* по расширенному множеству переменных ($\boldsymbol{x}$ и $\boldsymbol{z}$) совпадает с минимумом энергии E по переменным $\boldsymbol{x}$. Подробнее о сведении энергий высоких порядков к парно-сепарабельным написано в работе [49].

1.3.2. Приближённые алгоритмы

1.3.2.1. Шагающие алгоритмы

Шагающие алгоритмы (move-making algorithms) – это приближённые методы минимизации энергии, которые, начиная с некоторого начального приближения, итеративно переходят от одной разметки к другой. На каждой итерации рассматривается некоторое множество разметок, в которые можно перейти (совершить шаг), и выбирается разметка из данного множества, на которой энергия принимает минимальное (по текущему множеству) значение. Алгоритм останавливается, когда не может выполнить шаг, приводящий к уменьшению энергии. Интуитивно-очевидно, что алгоритм тем лучше, чем большее количество разметок покрывается на каждом

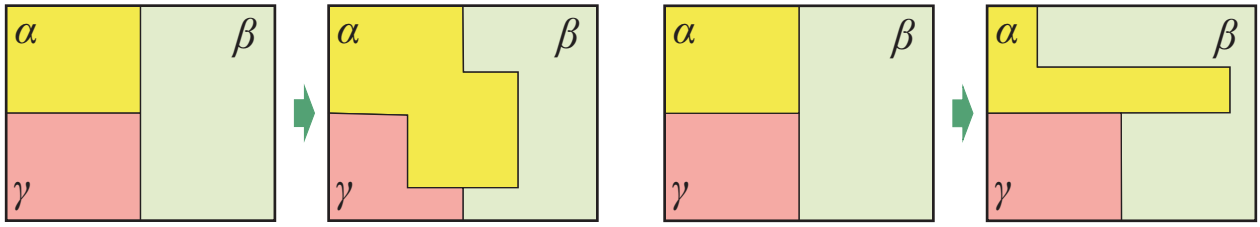
шаге. Современные алгоритмы данного класса, такие как α-расширение и $\alpha\beta$-замена [24], на каждой итерации выбирают шаг из экспоненциально большого количества возможных шагов и для многих реальных задач являются наиболее быстрыми [120]. Существуют алгоритмы, использующие и другие виды шагов: алгоритм range-moves [126] на каждом шаге минимизирует субмодулярные энергии от небинарных переменных, алгоритм fusion moves [87] использует приближённые алгоритмы минимизации бинарных несубмодулярных энергий. Недостатком алгоритмов данного класса является отсутствие гарантий оптимальности, а также ограниченная область применимости методов, позволяющих делать большие шаги.

Алгоритм ICM. Наиболее простым и хронологически первым шагающим методом является алгоритм ICM (Iterated Conditional Models) [19]. Шаги алгоритма ICM заключаются в выборе подмножетсва переменных $A \subseteq \mathcal{V}$, фиксации всех переменных, не входящих в множество A, и полному перебору по всем разметкам переменных, входящих в множество A. Сложность одного шага данного алгоритма экспоненциально мощности множества A, что делает алгоритм ICM неприменимым при больших A. Тем не менее, данный алгоритм применим для произвольных энергий и часто применяется, чтобы немного уточнить окончательное решение (например, в работе [111]), или в качестве эвристики, позволяющей получить прямое решение при решении двойственной задачи (например, в работе [96]).

Алгоритмы на основе разрезов графов. В 2001 г. Бойков и др. [24] предложили два шагающих алгоритма, основанных на итеративном применении алгоритма построения минимального разреза графа: α-расширение и $\alpha\beta$-замена. Каждый шаг обоих алгоритмов представляет собой задачу минимизации субмодулярной энергии бинарных переменных. В алгоритме α-расширение на каждом шаге каждая переменная из x_i, $i \in \mathcal{V}$ может либо получить выбранное значение $\alpha \in \mathcal{P}$, либо сохранить текущее значение. В алгоритме $\alpha\beta$-замена на каждом шаге все переменные, которым присвоены метки $\alpha, \beta \in \mathcal{P}$, могут получить значения α или β. Примеры шагов обоих алгоритмов приведены на рис. 1.5. Количество возможных шагов на каждой итерации обоих алгоритмов экспоненциально по количеству пикселей.

Приведем формальное описание каждого шага алгоритма α-расширение. Рассмотрим парно-сепарабельную энергию (1.2), заданную на графе $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Пусть есть текущая разметка $\boldsymbol{x}^0$ и выбрана «расширяемая» метка $\alpha \in \mathcal{P}$. Обозначим новую $|\mathcal{P}|$-значную разметку $\boldsymbol{x}^*$. Построим парно-сепарабельную энергию бинарных переменных

$$E_B(\boldsymbol{y}) = \sum_{i \in \mathcal{V}_B} \psi_i(y_i) + \sum_{\{i,j\} \in \mathcal{E}_B} \psi_{ij}(y_i, y_j)$$



(a) (b)

Рисунок 1.5.: Шаги алгоритмов α-расширение (a) и $\alpha\beta$-замена (b).

по энергии (1.2) следующим образом:

- Граф энергий $E_B(\boldsymbol{y})$ и $E(\boldsymbol{x})$ одинаков: $\mathcal{V}_B = \mathcal{V}$, $\mathcal{E}_B = \mathcal{E}$.
- Значение бинарной переменной $y_i = 0$ соответствует $x_i^* = x_i^0$ (значение переменной x_i не меняется); $y_i = 1$ соответствует $x_i^* = \alpha$.
- Унарные потенциалы: $\psi_i(0) = \theta_i(x_i^0)$, $\psi_i(1) = \theta_i(\alpha)$.
- Парные потенциалы: $\psi_{ij}(0,0) = \theta_{ij}(x_i^0, x_j^0)$, $\psi_{ij}(1,1) = \theta_{ij}(\alpha,\alpha)$, $\psi_{ij}(0,1) = \theta_{ij}(y_i^0, \alpha)$, $\psi_{ij}(1,0) = \theta_{ij}(\alpha, y_j^0)$.

Условием применимости алгоритма разреза графа является условие субмодулярности парных потенциалов (1.15). В терминах $|\mathcal{P}|$-значных переменных условие выглядит следующим образом:

$$\theta_{ij}(\alpha,\alpha) + \theta_{ij}(\beta,\gamma) \leqslant \theta_{ij}(\beta,\alpha) + \theta_{ij}(\alpha,\gamma), \tag{1.22}$$

где $\alpha, \beta, \gamma \in \mathcal{P}$. Часто предполагают, что $\theta_{ij}(\alpha,\alpha) = 0$ и $\theta_{ij}(\beta,\gamma) = \theta_{ij}(\gamma,\beta)$. В этом случае условие (1.22) становится неравенством треугольника, а все парные потенциалы ψ_{ij} – метриками.

Алгоритм α-расширение может быть существенно ускорен либо при помощи использования динамических разрезов графов [11], либо рассмотрения одновременно прямых и двойственных задач линейного программирования [72]. Алгоритм α-расширение может быть «совмещен» с алгоритмом $\alpha\beta$-замена для дальнейшего улучшения качества [108].

Алгоритм α-расширение можно обобщить и на случай потенциалов высоких порядков некоторых специальных видов: потенциалы Поттса высокого порядка ($\mathcal{P}^n$-Potts) [60] и их робастный вариант [61], потенциалы, штрафующие использование большого числа меток в решении [32], а также использующие информацию о частоте одновременного появления меток разных классов [80].

Современные работы по сравнению разных методов минимизации энергии [120, 54] показывают, что шагающие алгоритмы, основанные на разрезах графов, работают очень быстро, когда применимы.

1.3.2.2. Релаксационные алгоритмы

Одним из популярных подходов к решению задачи минимизации энергии (1.3) является *релаксация*. Данный подход заключается в формулировании задачи, как целочисленного программирования, после чего ограничения целочисленности отбрасываются, и решается непрерывная задача оптимизации.

Особенностью релаксационных методов является возможность предоставления так называемого *сертификата оптимальности*: если по решению непрерывной задачи удаётся построить решение дискретной задачи с таким же значением функционала, то построенное дискретное решение является глобально оптимальным. При решении непрерывных задач часто переходят к двойственным им. При таком переходе каждое значение двойственной функции является нижней оценкой на глобальный минимум, что, в свою очередь, позволяет говорить о построении наиболее точной нижней оценки среди определённого семейства оценок. Необходимым условием предоставления сертификата оптимальности является равенство нулю зазора[7] между предъявленным решением прямой дискретной задачей и наилучшей нижней оценкой на её глобальный минимум.

1.3.2.2.1. Линейная релаксация. Наиболее распространённой релаксацией является линейная релаксация, предложенная в 70-х годах М. И. Шлезингером [8], и позднее переоткрытая другими авторами [27, 130].

Рассмотрим задачу минимизации парно-сепарабельной энергии (1.2). Переформулируем её в виде эквивалентной задачи целочисленного линейного программирования (ЦЛП) с целевой функцией:

$$E_L(\boldsymbol{y}_L) = \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\theta_{ip}y_{ip} + \sum_{(i,j)\in\mathcal{E}}\sum_{p,q\in\mathcal{P}}\theta_{ij,pq}y_{ij,pq} \tag{1.23}$$

[7] Разность между значением глобального минимума непрерывной прямой задачи и значением глобального максимума двойственной задачи называется *зазором двойственности* (duality gap). Если непрерывная прямая задача является выпуклой, то обычно зазор двойственности равен 0 (в силу применимости выпуклого варианта теоремы Каруша-Куна-Таккера). *Зазором целочисленности* обычно называют разность между значением глобального минимума дискретной задачи и значением глобального минимума прямой непрерывной задачи. При решении прикладных задач особый интерес представляет сумма зазоров целочисленности и двойственности. Будем назвать эту величину просто *зазором*.

и множеством ограничений

$$\sum_{p\in\mathcal{P}} y_{ip} = 1, \quad \forall i \in \mathcal{V}, \tag{1.24}$$

$$\sum_{p\in\mathcal{P}} y_{ij,pq} = y_{jq}, \quad \forall\{i,j\} \in \mathcal{E}, \forall q \in \mathcal{P}, \tag{1.25}$$

$$\sum_{q\in\mathcal{P}} y_{ij,pq} = y_{ip}, \quad \forall\{i,j\} \in \mathcal{E}, \forall p \in \mathcal{P}, \tag{1.26}$$

$$y_{ij,pq}, y_{ip} \in \{0,1\}. \tag{1.27}$$

Ограничения (1.27) и (1.24) аналогично ограничениям (1.6) и (1.7) обеспечивают целочисленность и целостность (возможность единственным образом восстановить значения переменных $\boldsymbol{x}$ исходной задачи). Ограничения (1.25) и (1.26) обеспечивают согласованность унарных переменных y_{ip} и парных переменных $y_{ij,pq}$.

Уэйнрайт и др. [130] показали, что значение глобального минимума дискретной задачи (1.23)-(1.27) равно значению минимума линейного функционала (1.23) на некотором выпуклом многограннике, называемом *маргинальным многогранником* (marginal polytope). Охарактеризовать его полностью достаточно сложно, поскольку он имеет экспоненциальное количество граней. Одним из многогранников, содержащим в себе все целочисленные решения задачи (1.23)-(1.27), является многогранник, получаемый заменой ограничения целочисленности (1.27) на неотрицательность переменных $y_{ij,pq}$ и y_{ip}. Такой многогранник часто называют *локальным маргинальным многогранником* (local marginal polytope). Маргинальный и локальный маргинальный многогранники для энергий, определённых на графе $\mathcal{G}$, будет обозначать $\mathrm{Marginal}(\mathcal{G})$ и $\mathrm{Local}(\mathcal{G})$, соответственно.

Определение 5. *Стандартной линейной релаксацией задачи минимизации дискретной парносепарабельной энергии* (1.2) *назовём задачу минимизации линейного функционала* (1.23) *по переменным* y_{ip}, $y_{ij,pq}$, *принимающим действительные неотрицательные значения, при ограничениях* (1.24), (1.25), *и* (1.26).

В некоторых ситуациях зазор между решением дискретной задачи (1.23)-(1.27) и её стандартной релаксацией может быть достаточно большим. В последние годы был разработан целый ряд методов уточнения стандартной релаксации, так или иначе основанных на добавлении дополнительных линейных ограничений: [133, 70, 116, 17]. Тем не менее, во многих практически важных случаях зазор либо отсутствует, либо мал, что делает стандартную линейную релаксацию популярным инструментом при решении прикладных задач. Можно показать, что для ряда важных частных случаев стандартная линейная релаксация является точной.

Утверждение 2. *Если граф $\mathcal{G}$, на котором задана парно-сепарабельная энергия $E(\boldsymbol{x})$, не содержит циклов, то значение глобального минимума энергии $E(\boldsymbol{x})$ в точности совпадает со значением минимума стандартной линейной релаксации задачи минимизации энергии $E(\boldsymbol{x})$.*

Утверждение 3. *Если парно-сепарабельная энергия бинарных переменных $E(\boldsymbol{x})$, заданная на произвольном графе $\mathcal{G}$, субмодулярна, то значение глобального минимума энергии $E(\boldsymbol{x})$ в точности совпадает со значением минимума стандартной линейной релаксации задачи минимизации энергии $E(\boldsymbol{x})$.*

Доказательства утв. 2 и 3 представлены в работах [129] и [67], соответственно.

Стандартная линейная релаксация описывается полиномиальным множеством ограничений. Тем не менее, во многих реальных ситуациях в возникающей задаче линейного программирования очень много переменных, что делает невозможным применение существующих методов общего назначения (например, симплекс-метода). Существует целый ряд специализированных методов для решения именно такой задачи линейного программирования. Большинство методов строят двойственную задачу тем или иным способом, после чего решается оптимизационная задача в пространстве двойственных переменных. Все такие методы можно разделить на два класса:

- двойственная задача выписывается в явном виде;
- двойственная задача строится при помощи двойственного разложения (dual decompostion), что не позволяет выписать её в явном виде.

Явный вид двойственной задачи. Построим двойственную задачу в явном виде, предложенном М. И. Шлезингером [8, 132]. Рассмотрим задачу минимизации парно-сепарабельной энергии (1.2). Добавим и вычтем новые слагаемые следующим образом:

$$E(\boldsymbol{x}) = \sum_{i\in\mathcal{V}} \left(\theta_i(x_i) + \sum_{j:\{i,j\}\in\mathcal{E}} m_{ji}(x_i) \right) + \sum_{\{i,j\}\in\mathcal{E}} \left(\theta_{ij}(x_i,x_j) - m_{ji}(x_i) - m_{ij}(x_j)\right). \tag{1.28}$$

Здесь $m_{ji}(x_i)$ – добавленные слагаемые, которые часто в силу сходства с $\mu_{i\to j}(x_j)$ (1.8) называют сообщениями. Группируя слагаемые согласно выражению (1.28), получим следующую нижнюю оценку на энергию $E(\boldsymbol{x})$:

$$\begin{aligned} D\left(\{m_{ji}(x_i),\ m_{ij}(x_j)\}_{\{i,j\}\in\mathcal{E}}\right) = &\sum_{i\in\mathcal{V}} \min_{x_i\in\mathcal{P}} \left(\theta_i(x_i) + \sum_{j:\{i,j\}\in\mathcal{E}} m_{ji}(x_i) \right) \\ &+ \sum_{\{i,j\}\in\mathcal{E}} \min_{x_i,x_j\in\mathcal{P}} \left(\theta_{ij}(x_i,x_j) - m_{ji}(x_i) - m_{ij}(x_j)\right). \end{aligned} \tag{1.29}$$

Задачу максимизации двойственной функции (1.29) можно переписать в виде задачи линейного программирования, двойственной к стандартной линейной релаксации исходной задачи:

$$\begin{aligned}\max_{m_{ji}(x_i),\, m_{ij}(x_j),\, h_i,\, h_{ij}} \quad & \sum_{i\in\mathcal{V}} h_i + \sum_{\{i,j\}\in\mathcal{E}} h_{ij} \\ \text{s.t.} \quad & h_i \geqslant \theta_i(x_i) + \sum_{j:\{i,j\}\in\mathcal{E}} m_{ji}(x_i), \quad \forall i \in \mathcal{V},\ x_i \in \mathcal{P}, \\ & h_{ij} \geqslant \theta_{ij}(x_i, x_j) - m_{ji}(x_i) - m_{ij}(x_j), \quad \forall \{i,j\} \in \mathcal{E},\ x_i, x_j \in \mathcal{P},\end{aligned}$$

Запись двойственной задачи оптимизации в явном виде позволяет получать схемы блочно-координатной оптимизации, гарантирующие монотонное неубывание нижней оценки на каждом шаге. Алгоритмами данного класса являются алгоритмы MSD [8, 5, 132], а также различные варианты алгоритма MPLP [42, 117].

Общим инструментом выписывания двойственных задач в явном виде является двойственность по Фенхелю. Зак и др. [139] используют двойственность по Фенхелю для построения двойственных функций, которые можно сглаживать, что, в свою очередь, позволяет применять гладкие методы оптимизации для решения двойственных задач. Похожая идея была применена Савчинским и др. [106] для ситуации, когда двойственную функцию в явном виде выписать не удаётся.

Парно-сепарабельные бинарные энергии. Особняком стоит ситуация стандартной линейной релаксации для случая парно-сепарабельных энергий бинарных переменных (в том числе, несубмодулярных). В этом случае стандартную линейную релаксации можно эффективно найти при помощи построения разреза специального графа. В литературе эта техника обычно называется алгоритмом QPBO.

Для каждой переменной $x_i \in \{0, 1\}$, $i \in \mathcal{V}$ введём дополнительную переменную $\bar{x}_i = 1 - x_i$. Используя новые переменные, исходную энергию можно записать как субмодулярную функцию[8]. Если после этого отбросить ограничения $x_i + \bar{x}_i = 1$, $i \in \mathcal{V}$, то записанную функцию можно будет эффективно минимизировать при помощи построения разреза графа (см. раздел 1.3.1.2). Можно доказать несколько свойств такой релаксации:

- величина минимального разреза построенного графа равна значению минимума стандартной линейной релаксации (лемма 1 и теоремы 6 и 10 из работы [22]);

[8] Пусть, без ограничения общности, все значения и унарных, и парных потенциалов неотрицательны. Тогда унарные потенциалы можно записать, например, так: $\theta_i(x_i) = \theta_i(0)\bar{x}_i + \theta_i(1)x_i$. Парные так: $\theta_{ij}(x_i, x_j) = \theta_{ij}(0,0)(1-x_i)\bar{x}_j + \theta_{ij}(0,1)(1-x_i)x_j + \theta_{ij}(1,0)x_i(1-x_j) + \theta_{ij}(1,1)x_i(1-\bar{x}_j)$. При такой записи всех потенциалов энергия, как функция расширенного множества переменных, будет субмодулярна.

- если для некоторого множества вершин $\mathcal{W} \subseteq \mathcal{V}$ для разметки, построенной по разрезу графа, выполнено условие $x_i + \bar{x}_i = 1$, то существует оптимальная разметка $\boldsymbol{x}^*$ исходной энергии, в которой $x_i^* = x_i$, $i \in W$ (частичная оптимальность) [66, 47];
- если для произвольной разметки исходных переменных $\tilde{\boldsymbol{x}}$ положить $\hat{x}_i = x_i$, $i \in \mathcal{W}$, и $\hat{x}_i = \tilde{x}_i$, $i \notin \mathcal{W}$, то энергия при этом не увеличится $E(\hat{\boldsymbol{x}}) \leqslant E(\tilde{\boldsymbol{x}})$ (weak autarky) [103, 66, 47].

Обычно говорят, что алгоритм QPBO строит частичную разметку переменных $\boldsymbol{x}$ – присваивает каждой переменной x_i, $i \in \mathcal{V}$ значение из множества $\{0, 1, \varnothing\}$, где элемент $\varnothing$ соответствует отказу от разметки. В стандартной линейной релаксации, решаемой алгоритмом QPBO, при этом выполнено свойство *полуцелочисленности* (half-integrality): существует решение, в котором все переменные принимают значения из множества $\{0, 0.5, 1\}$.

Существуют обобщения алгоритма QPBO на случай энергий с небинарными переменными [74, 63], а также методы, использующие свойство частичной оптимальности для повышения производительности [11] и улучшения качества решений [87, 107] .

1.3.2.2.2. Двойственная декомпозиция. Комодакис и др. [73] сформулировали идею декомпозиции графа задачи минимизации энергии на подзадачи с ациклическими графами [130, 65] через релаксацию Лагранжа. Такой подход является более простым для понимания (чем подход, основанный на передаче сообщений [130, 65]) и обладает большей гибкостью.

Рассмотрим задачу минимизации парно-сепарабельной энергии (1.2), определённой на произвольном графе $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Разобьём энергию на слагаемые следующим образом:

$$E(\boldsymbol{x}) = \sum_{S \in \mathcal{S}} E^S(\boldsymbol{x}_S) = \sum_{S \in \mathcal{S}} \left(\sum_{i \in \mathcal{V}^S} \theta_i^S(x_i) + \sum_{\{i,j\} \in \mathcal{E}^S} \theta_{ij}^S(x_i, x_j) \right).$$

Здесь символ S индексирует слагаемые (*подэнергии*), $\mathcal{S}$ – конечное множество индексов слагаемых. Подэнергии E^S являются энергиями, заданными на графах $\mathcal{G}^S = (\mathcal{V}^S, \mathcal{E}^S)$, где $\mathcal{V}^S \subseteq \mathcal{V}$ и $\mathcal{E}^S \subseteq \mathcal{E}$, а также $\bigcup_{S \in \mathcal{S}} \mathcal{V}^S = \mathcal{V}$, $\bigcup_{S \in \mathcal{S}} \mathcal{E}^S = \mathcal{E}$. Символ $\boldsymbol{x}_S$ обозначает переменные, индексы которых входят в множество $\mathcal{V}^S$ (сокращение от $\boldsymbol{x}_{\mathcal{V}^S}$). Потенциалы каждой из подэнергий в сумме составляют потенциалы энергий:

$$\theta_i(x_i) = \sum_{S \in \mathcal{S}\colon\, i \in \mathcal{V}^S} \theta_i^S(x_i), \quad \forall i \in \mathcal{V};$$

$$\theta_{ij}(x_i, x_j) = \sum_{S \in \mathcal{S}\colon\, \{i,j\} \in \mathcal{E}^S} \theta_{ij}^S(x_i, x_j), \quad \forall \{i,j\} \in \mathcal{E}.$$

Решение задачи минимизации подэнергии $E^S(\boldsymbol{x}^S)$[9] (подзадачи) может оказаться существенно проще, чем решение задачи минимизации исходной энергии $E(\boldsymbol{x})$. Например, если граф $\mathcal{G}^S$ подзадачи S не содержит циклов, то эту подзадачу можно эффективно решить точно при помощи алгоритмов передачи сообщений (см. раздел 1.3.1.1.1). Предположим, что разбиение на подэнергии выбрано так, что все графы $\mathcal{G}^S$, $S \in \mathcal{S}$, не содержат циклов, а значит задачи $\min_{\boldsymbol{x}^S} E^S(\boldsymbol{x}^S)$ можно решить точно. Решая все подзадачи независимо, можно получить нижнюю оценку на глобальный минимум исходной задачи:

$$\min_{\boldsymbol{x}} E(\boldsymbol{x}) \geqslant \sum_{S \in \mathcal{S}} \min_{\boldsymbol{x}^S} E^S(\boldsymbol{x}^S).$$

Данную нижнюю оценку можно существенно уточнить, обеспечивая согласование переменных подзадач, при помощи релаксации Лагранжа. Сформулируем исходную задачу минимизации энергии как задачу условной оптимизации, относительно индикаторных переменных подзадач y^S_{ip}, $y^S_{ij,pq}$ (аналогично задаче ЦЛП 1.23-1.27):

$$\min_{\boldsymbol{y}^S \in \mathrm{Local}(\mathcal{G}^S),\, \boldsymbol{y}^*} \quad \sum_{S \in \mathcal{S}} E^S_L(\boldsymbol{y}^S), \tag{1.30}$$

$$\text{s.t.} \quad y^S_{ip} = y^*_{ip}, \quad \forall i \in \mathcal{V}, \forall p \in \mathcal{P}, \forall S \in \mathcal{S} : i \in \mathcal{V}^S, \tag{1.31}$$

$$y^S_{ij,pq} = y^*_{ij,pq}, \quad \forall \{i,j\} \in \mathcal{E}, \forall p,q \in \mathcal{P}, \forall S \in \mathcal{S} : \{i,j\} \in \mathcal{E}^S. \tag{1.32}$$

Здесь $\boldsymbol{y}^S = \left\{ \{y_{ip}\}^{p \in \mathcal{P}}_{i \in \mathcal{V}^S}, \{y_{ij,pq}\}^{p,q \in \mathcal{P}}_{\{i,j\} \in \mathcal{E}^S} \right\}$ – переменные, входящие в состав подзадач, $\boldsymbol{y}^*$ – добавленные переменные, обеспечивающие согласование подзадач, $E^S_L(\boldsymbol{y}^S)$ – целевая функция стандартной линейной релаксации (1.23) подзадачи S. Поскольку стандартная линейная релаксация на дереве точна (утв. 2), можно писать, что переменные $\boldsymbol{y}^S$ лежат в многограннике $\mathrm{Local}(\mathcal{G}^S)$ (а не принадлежат дискретному множеству с ограничениями вида (1.24), (1.25), (1.26), (1.27)). Запишем лагранжиан задачи условной оптимизации (1.30)-(1.32):

$$L(\boldsymbol{y}^S, \boldsymbol{y}^*, \boldsymbol{\lambda}^S) = \sum_{S \in \mathcal{S}} E^S_L(\boldsymbol{y}^S) + \sum_{S \in \mathcal{S}} \left\langle \boldsymbol{\lambda}^S, \boldsymbol{y}^S - \boldsymbol{y}^*_S \right\rangle, \tag{1.33}$$

где $\boldsymbol{y}^*_S$ – часть переменных вектора $\boldsymbol{y}^*$, относящиеся ко множествам $\mathcal{V}^S$ и $\mathcal{E}^S$; $\boldsymbol{\lambda}^S$ – множители Лагранжа для ограничений (1.31) и (1.32), относящиеся к подзадаче S. Символы $\langle \cdot, \cdot \rangle$ соответствуют скалярному произведению. Используя лагранжиан, получим нижнюю оценку на решение

[9] Заметим, что в символы $\boldsymbol{x}^S$ и $\boldsymbol{x}_S$ вкладывается несколько разный смысл. Символ $\boldsymbol{x}_S$ означает проекцию переменных $\boldsymbol{x}$ на множество $\mathcal{V}^S \subseteq \mathcal{V}$, а $\boldsymbol{x}^S$ – независимые переменные, соответствующие такому же множеству. Принципиальная разница состоит в том, что если используется запись $\boldsymbol{x}_S$, то подразумевается, что в данном выражении переменные для разных множеств $\mathcal{V}^{S_1}$ и $\mathcal{V}^{S_2}$ согласованы: $i \in \mathcal{V}^{S_1} \cap \mathcal{V}^{S_2} \;\Rightarrow\; x_{S_1,i} = x_{S_2,i}$.

оптимизационной задачи (1.30)-(1.32):

$$\min_{\substack{\boldsymbol{y}^S,\, \boldsymbol{y}^*\in(1.31),(1.32):\\ \boldsymbol{y}^S\in\mathrm{Local}(\mathcal{G}^S)}} \sum_{S\in\mathcal{S}} E_L^S(\boldsymbol{y}^S) = \min_{\boldsymbol{y}^S\in\mathrm{Local}(\mathcal{G}^S),\, \boldsymbol{y}^*} \max_{\boldsymbol{\lambda}^S} L(\boldsymbol{y}^S, \boldsymbol{y}^*, \boldsymbol{\lambda}^S)$$

$$\geqslant \max_{\boldsymbol{\lambda}^S} \min_{\boldsymbol{y}^S\in\mathrm{Local}(\mathcal{G}^S),\, \boldsymbol{y}^*} L(\boldsymbol{y}^S, \boldsymbol{y}^*, \boldsymbol{\lambda}^S) = \max_{\boldsymbol{\lambda}^S} D(\boldsymbol{\lambda}^S). \quad (1.34)$$

Здесь $D(\boldsymbol{\lambda}^S)$ – двойственная функция.

При минимизации лагранжиана (1.33) по переменным $\boldsymbol{y}^*$ на эти переменные нет никаких ограничений. Поскольку эти переменные входят линейно, то везде, где коэффициенты перед ними не равны нулю, значение минимума лагранжиана не ограничено снизу. Из этого следует, что переменные $\boldsymbol{\lambda}^S$, по которым проводится максимизация, никогда не примут значений, при которых возникнут ненулевые коэффициенты. Данное рассуждение позволяет записать задачу максимизации нижней оценки (1.34) в виде задачи условной оптимизации, в которой переменные $\boldsymbol{y}^*$ не присутствуют:

$$\max_{\boldsymbol{\lambda}^S} \quad \sum_{S\in\mathcal{S}} \min_{\boldsymbol{y}^S\in\mathrm{Local}(\mathcal{G}^S)} \left(E_L^S(\boldsymbol{y}^S) + \left\langle \boldsymbol{\lambda}^S, \boldsymbol{y}^S \right\rangle\right), \quad (1.35)$$

$$\text{s.t.} \quad \sum_{S\in\mathcal{S}:\, i\in\mathcal{V}^S} \lambda_{ip}^S = 0, \quad \forall i\in\mathcal{V}, \forall p\in\mathcal{P}, \quad (1.36)$$

$$\sum_{S\in\mathcal{S}:\, \{i,j\}\in\mathcal{E}^S} \lambda_{ij,pq}^S = 0, \quad \forall\{i,j\}\in\mathcal{E}, \forall p,q\in\mathcal{P}. \quad (1.37)$$

Для вычисления целевой функции (1.35) при произвольном значении переменных $\boldsymbol{\lambda}^S$ необходимо решать задачи оптимизации по переменным $\boldsymbol{y}^S$ вида

$$\min_{\boldsymbol{y}^S\in\mathrm{Local}(\mathcal{G}^S)} \left(E_L^S(\boldsymbol{y}^S) + \left\langle \boldsymbol{\lambda}^S, \boldsymbol{y}^S \right\rangle\right). \quad (1.38)$$

Эти задачи являются задачами минимизации парно-сепарабельной энергии на графе без циклов, а значит могут быть эффективно решены при помощи алгоритма передачи сообщений.

Задача (1.35)-(1.37) является задачей максимизации вогнутой кусочно-линейной функции (1.35) при ограничениях (1.36)-(1.37). Данную задачу можно решать методом проекции субградиента[10]. Используя решения подзадач (1.38), субградиент можно вычислить следующим

[10] *Субградиент* является обобщением понятия градиента для случая негладких функций. Для вогнутой функции $f(\boldsymbol{\lambda})$ субградиент в точке $\boldsymbol{\lambda}_0$ – это такой вектор $\boldsymbol{g}$, при котором для любых $\boldsymbol{\lambda}$ выполнено

$$f(\boldsymbol{\lambda}) - f(\boldsymbol{\lambda}_0) \geqslant \langle \boldsymbol{g}, \boldsymbol{\lambda} - \boldsymbol{\lambda}_0 \rangle.$$

Если в точке $\boldsymbol{\lambda}_0$ функция $f(\boldsymbol{\lambda})$ дифференцируема, то субградиент определён единственным образом и совпадает с градиентом. Если же функция $f(\boldsymbol{\lambda})$ в точке $\boldsymbol{\lambda}_0$ не дифференцируема, то субградиент не единственен.

образом:

$$\partial D(\boldsymbol{\lambda}^S)_{ip} = y^S_{ip}, \qquad \forall i \in \mathcal{V}, \forall p \in \mathcal{P}, \forall S \in \mathcal{S} : \; i \in \mathcal{V}^S;$$
$$\partial D(\boldsymbol{\lambda}^S)_{ij,pq} = y^S_{ij,pq}, \qquad \forall \{i,j\} \in \mathcal{E}, \forall p,q \in \mathcal{P}, \forall S \in \mathcal{S} : \; \{i,j\} \in \mathcal{E}^S.$$

Проекцию субградиента на множество, заданное ограничениями (1.36) и (1.37), можно вычислить следующим образом:

$$\Delta\lambda^S_{ip} = y^S_{ip} - \frac{\sum_{S' \in \mathcal{S}: i \in \mathcal{V}^{S'}} y^{S'}_{ip}}{|\{S' \in \mathcal{S} : i \in \mathcal{V}^{S'}\}|}, \qquad \forall i \in \mathcal{V}, \forall p \in \mathcal{P}, \forall S \in \mathcal{S} : \; i \in \mathcal{V}^S;$$
$$\Delta\lambda^S_{ij,pq} = y^S_{ij,pq} - \frac{\sum_{S' \in \mathcal{S}: \{i,j\} \in \mathcal{E}^{S'}} y^{S'}_{ij,pq}}{|\{S' \in \mathcal{S} : \{i,j\} \in \mathcal{E}^{S'}\}|}, \qquad \forall \{i,j\} \in \mathcal{E}, \forall p,q \in \mathcal{P}, \forall S \in \mathcal{S} : \; \{i,j\} \in \mathcal{E}^S.$$

Заметим, что количество двойственных переменных $\boldsymbol{\lambda}^S$ существенно зависит от способа разбиения задачи на деревья. Если каждое ребро исходного графа $\mathcal{G}$ покрывается одним и только одним подграфом $\mathcal{G}^S$, то ограничения (1.37), а значит и переменные $\lambda^S_{ij,pq}$, не нужны. Если каждая вершина покрывается только двумя подграфами, то ограничения (1.36) можно заменить на более простые: $x^{S_1}_{ip} = x^{S_2}_{ip}$. Примером такой ситуации может случить граф $\mathcal{G}$ вида 4-х связная решётка, разбитый на два подграфа, содержащих только вертикальные или только горизонтальные рёбра.

Комодакис и др. [73] доказали следующее утверждение о наилучшей нижней оценке решения задачи минимизации энергии $E(\boldsymbol{x})$, построенной при помощи решения непрерывной задачи (1.35)-(1.37):

Утверждение 4. *Если графы $\mathcal{G}^S$ всех подзадач $S \in \mathcal{S}$ не содержат циклов, то значение глобального максимума задачи* (1.35)-(1.37) *в точности совпадает со значением глобального минимума стандартной линейной релаксации исходной задачи.*

Это утверждение означает, что оптимальное значение нижней оценки, которую можно получить при помощи двойственной декомпозиции с разбиением графа задачи на ациклические подграфы (деревья), не зависит от конкретного способа разбиения. Тем не менее, конкретный способ разбиения на деревья может существенно влиять на скорость сходимости метода.

В рамках данной работы метод минимизации парно-сепарабельной энергии при помощи декомпозиции графа задачи на деревья и дальнейшего согласования при помощи релаксации Лагранжа будем называть DD TRW (dual decomposition tree reweighted message passing).

В работах [70, 16, 136, 137, 53] показано, что метод двойственной декомпозиции может получать нижние оценки, более точные, чем стандартная линейная релаксация. Для этого надо использовать подзадачи, которые можно решить точно и при этом стандартная линейная релаксация на них не точна. Примерами таких подзадач могут служить задачи с графами малой ширины [82, 129] и энергии моделей Изинга для планарных графов [109].

1.3.2.2.3. Двойственная декомпозиция для случая потенциалов высоких порядков. Комодакис и др. [71] исследовали возможность применения подхода двойственной декомпозиции для минимизации энергий с потенциалами высоких порядков (1.1).

Наиболее простым способом использования подхода двойственной декомпозиции является введение подзадач на каждый из потенциалов высоких порядков. Аналогично разделу 1.3.2.2.2 можно получить нижнюю оценку на глобальный минимум энергии:

$$\begin{aligned}
\min_{\boldsymbol{y}} E_(\boldsymbol{y}) &= \min_{\substack{\{\boldsymbol{y}^C\}_{C\in\mathcal{C}},\,\boldsymbol{y}^*:\\ \boldsymbol{y}^*_C=\boldsymbol{y}^C}} \sum_{C\in\mathcal{C}} \left(\theta_C(\boldsymbol{y}^C) + \sum_{i\in C}\sum_{p\in\mathcal{P}} \theta_{ip} y^C_{ip}/|\mathcal{C}(i)| \right) \\
&= \min_{\{\boldsymbol{y}^C\},\boldsymbol{y}^*} \max_{\{\lambda_{ip,C}\}} \left(\sum_{C\in\mathcal{C}} \left(\theta_C(\boldsymbol{y}^C) + \sum_{i\in C}\sum_{p\in\mathcal{P}} \theta_{ip} y^C_{ip}/|\mathcal{C}(i)| \right) + \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\sum_{C\in\mathcal{C}(i)} \lambda_{ip,C}(y^C_{ip} - y^*_{ip}) \right) \\
&\geqslant \max_{\{\lambda_{ip,C}\}} \min_{\{\boldsymbol{y}^C\},\boldsymbol{y}^*} \left(\sum_{C\in\mathcal{C}} \left(\theta_C(\boldsymbol{y}^C) + \sum_{i\in C}\sum_{p\in\mathcal{P}} y^C_{ip}\left(\theta_{ip}/|\mathcal{C}(i)| + \lambda_{i,C}\right) \right) - \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\sum_{C\in\mathcal{C}(i)} \lambda_{ip,C} y^*_{ip} \right) \\
&= \max_{\substack{\{\lambda_{ip,C}\}:\\ \sum_{C\in\mathcal{C}(i)}\lambda_{ip,C}=0}} \min_{\{\boldsymbol{y}^C\}} \left(\theta_C(\boldsymbol{y}^C) + \sum_{i\in C}\sum_{p\in\mathcal{P}} y^C_{ip}\left(\theta_{ip}/|\mathcal{C}(i)| + \lambda_{i,C}\right) \right) \\
&= \max_{\substack{\{\lambda_{ip,C}\}:\\ \sum_{C\in\mathcal{C}(i)}\lambda_{ip,C}=0}} D(\boldsymbol{\lambda})
\end{aligned} \tag{1.39}$$

Необходимым условием работы с нижней оценкой энергии (1.39) является возможность эффективно решать задачи минимизации, возникающие при вычислении нижней оценки $D(\boldsymbol{\lambda})$ при фиксированных множителях Лагранжа $\boldsymbol{\lambda}$. Данные задачи имеют вид суммы одного потенциала высокого порядка и унарных потенциалов:

$$\min_{\boldsymbol{x}_C\in\mathcal{X}_C} \left(\theta_C(\boldsymbol{x}_C) + \sum_{i\in C} \theta_i(x_i) \right). \tag{1.40}$$

Если порядок потенциалов небольшой, то задачи (1.40) можно решать при помощи полного перебора.

При порядках потенциалов, когда перебор становится вычислительно неприемлем, возникает проблема хранения потенциалов в памяти. Необходимо некоторое компактное представление, требующее значительно меньше памяти, чем задание таблицей. Одним из способов такого компактного задания является задание значения по умолчанию и небольшого числа значений в выделенных конфигурациях. Такие потенциалы были независимо рассмотрены в работах [104, 71] и названы разреженными (sparse) и заданным шаблонами (pattern-based).

Будем говорить, что потенциал θ_C *задан шаблонами*, если он принимает ненулевые значения только на заранее выделенном множестве конфигураций переменных $\mathcal{D}_C \subseteq \mathcal{X}_C$, а в осталь-

ных случаях принимает значение по умолчанию – 0:

$$\theta_C(\boldsymbol{x}_C) = \begin{cases} \theta_{C,\boldsymbol{d}}, & \text{если } \boldsymbol{x}_C = \boldsymbol{d},\ \boldsymbol{d} \in \mathcal{D}_C, \\ 0, & \text{иначе.} \end{cases}$$

Шаблонами может быть задан произвольный потенциал θ_C, но при этом множество $\mathcal{D}_C$ может быть очень большого размера. Будем говорить, что потенциал θ_C является *разреженным*, если он задан шаблонами, мощность множества $\mathcal{D}_C$ много меньше числа всех возможных конфигураций $|\mathcal{P}|^C$, и все коэффициенты $\theta_{C,\boldsymbol{d}}$ отрицательны.

Примером разреженных потенциалов второго порядка являются потенциалы Поттса (Potts): $\theta_{\{i,j\}}(x_i, x_j) = -[x_i = x_j]$. Примером разреженных потенциалов высоких порядков являются потенциалы Поттса высокого порядка ($\mathcal{P}^n$-Potts), введённые в работе [60]:

$$\theta_C(\boldsymbol{x}_C) = \begin{cases} -1, & \text{если все переменные } x_i,\ i \in C, \text{ принимают одинаковые значения,} \\ 0, & \text{иначе.} \end{cases}$$

Если потенциалы высокого порядка разреженные, то задачи (1.40) можно решить эффективно. В этом случае минимум может достигаться либо на одной из конфигураций множества выделенных конфигураций $\mathcal{D}_C$ (обычно $|\mathcal{D}_C| \ll |\mathcal{X}_C|$) или на конфигурации, доставляющей минимум унарных потенциалов: $x_i = \arg\min_{x_i \in \mathcal{P}} \theta_i(x_i)$, $i \in C$. Перебор всех этих конфигураций и приводит к решению задачи (1.40).

Аналогично задаче (1.35)-(1.37) задача максимизации нижней оценки (1.39) является кусочно-линейной и вогнутой, а значит её можно решить при помощи методов проекции субградиента.

Заметим, что нижнюю оценку (1.39) можно комбинировать с разложением парно-сепарабельных энергий на деревья, а именно объединять потенциалы порядка 2, если таковые присутствуют в энергии, в группы, так чтобы графы не образовывали циклов. В рамках данной работы будем называть этот метод CWD (clique-wise decomposition).

В работе [71] предложен способ объединения нескольких разреженных потенциалов высоких порядков в группу, допускающий эффективное решение задачи аналогичной (1.40) (алгоритм PatB).

1.3.2.2.4. Альтернативные виды релаксаций. Стандартная линейная релаксация является не единственной непрерывной релаксацией задачи минимизации энергии (1.5)-(1.7). Существуют и более точные линейные релаксации [115, 70, 116, 17], и другие виды релаксаций: квадратичная релаксация [101], коническая релаксация [78], полуопределённая (semidefinite, SDP) релаксация [131].

В семействах линейных релаксаций, более точных чем стандартная релаксация, обычно делается попытка построить приближение маргинального многогранника, более точное чем локальный маргинальный многогранник. Примером такого приближения может служить многогранник, описываемый всеми циклическими ограничениями (cycle inequalities) [115, 70]. Несмотря на, вообще говоря, более точные нижние оценки, методы этой группы редко используются на практике, поскольку являются достаточно медленными.

Квадратичная релаксация в формулировке из работы [101] всегда точна (зазор равен 0), но часто является не выпуклой. Для её использования приходится строить дополнительную нижнюю оценку (QP-RL), которая, в свою очередь, уже является выпуклой. Коническая релаксация второго порядка [78] (SOCP-MS) является выпуклой оптимизационной задачей, но в ней возможен ненулевой зазор. Кумар и др. [79] провели теоретическое сравнение трёх подходов (стандартная линейная релаксация, QP-RL, SOCP-MS) и показали, что стандартная линейная релаксация предоставляет нижнюю оценку строго лучше нижних оценок двух других подходов. Также Кумар и др. [79] предложили способ построения более точного семейства оценок на основе конического программирования, но для этого семейства не известно эффективных алгоритмов поиска наилучшей оценки. SDP-релаксация была предложена лишь в 2013 г., и её сравнение с другими видами релаксаций ещё не проведено.

2. Субмодулярная релаксация

В этой главе описан подход к задаче минимизации энергии, основанный на применении релаксации Лагранжа к ограничениям целостности: субмодулярная релаксация (submodular relaxation, SMR). Раздел 2.1 описывает частный случай SMR для парно-сепарабельных ассоциативных энергий: субмодулярное разложение (submodular decomposition, SMD), предложенное в работах [128, 97]. Раздел 2.2 описывает расширение SMD на случай энергий с потенциалами высоких порядков [96], раздел 2.3 – на случай парных потенциалов произвольного вида [127]. Раздел 2.4 описывает способ учёта глобальных линейных ограничений в рамках SMR.

2.1. Парно-сепарабельные ассоциативные энергии

Рассмотрим задачу минимизации парно-сепарабельной энергии в индикаторной нотации:

$$\min_{\boldsymbol{y}} \quad \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\theta_{ip}y_{ip} + \sum_{\{i,j\}\in\mathcal{E}}\sum_{p,q\in\mathcal{P}}\theta_{ij,pq}y_{ip}y_{jq}, \tag{2.1}$$

$$\text{s.t.} \quad \sum_{p\in\mathcal{P}} y_{ip} = 1, \quad \forall i \in \mathcal{V}, \tag{2.2}$$

$$y_{ip} \in \{0,1\}, \quad \forall i \in \mathcal{V}, p \in \mathcal{P}, \tag{2.3}$$

Ограничения (2.2) гарантируют, что каждая переменная i исходной задачи принимает одно и только одно значение из множества меток $\mathcal{P}$ (ограничения целостности), ограничения (2.3) – делают индикаторные переменные y_{ip} бинарными (ограничения целочисленности).

Пусть все парные потенциалы θ_{ij} принадлежат классу *ассоциативных* потенциалов [123]:

$$\theta_{ij}(p,q) = -C_{ij,p}[p=q] = \begin{cases} -C_{ij,p}, & p = q, \\ 0, & p \neq q, \end{cases} \tag{2.4}$$

где $C_{ij,p}$ – неотрицательные константы. Стоит отметить, что если значения $C_{ij,p}$ в случае равенства меток p и q не зависят от метки p ($C_{ij,p} = C_{ij}$), то ассоциативные потенциалы (2.4) становятся потенциалами Поттса (отличаются на константу от стандартной записи из работы [24]).

В случае ассоциативных парных потенциалов (2.4) ненулевые коэффициенты в целевой функции (2.1) могут присутствовать только при слагаемых, отвечающих унарным потенциалам, а также парным потенциалам вида $y_{ip}y_{jp}$, $p \in \mathcal{P}$. В этом случае индикаторы y_{ip}, относящиеся к разным меткам, связаны только ограничениями целостности (2.2). Таким образом, если отбросить ограничения целостности, то целевая функция распадается на $|\mathcal{P}|$ групп слагаемых (подзадач), каждая из которых содержит только переменные, относящиеся к определённой метке p.

Формально, целевая функция (2.1) может быть представлена в виде суммы слагаемых $E_I^p(\boldsymbol{y}_p)$, каждое их которых зависит только от переменных $\boldsymbol{y}_p = \{y_{ip} \mid i \in \mathcal{V}\}$, относящихся к одной метке $p \in \mathcal{P}$. Слагаемые можно записать следующим способом:

$$E_I^p(\boldsymbol{y}_p) = \sum_{j \in \mathcal{V}} \theta_{jp} y_{jp} - \sum_{\{i,j\} \in \mathcal{E}} C_{ij,p} y_{ip} y_{jp}. \tag{2.5}$$

Используя разложение целевой функции (2.1) на слагаемые (2.5), можно получить нижнюю оценку на глобальный минимум оптимизационной задачи (2.1)-(2.3):

$$\min_{\boldsymbol{y} \in (2.2),(2.3)} E_I(\boldsymbol{y}) = \min_{\boldsymbol{y} \in (2.2),(2.3)} \sum_{p \in \mathcal{P}} E_I^p(\boldsymbol{y}_p) \geqslant \sum_{p \in \mathcal{P}} \min_{\boldsymbol{y}_p \in \{0,1\}^{|\mathcal{V}|}} E_I^p(\boldsymbol{y}_p). \tag{2.6}$$

Функции E_I^p – парно-сепарабельные функции переменных $\boldsymbol{y}_p$. Согласно утв. 1 (т.к. $C_{ij,p} \geqslant 0$) эти функции субмодулярны, а значит могут быть эффективно минимизированы при помощи применения алгоритмов разрезов графов (см. раздел 1.3.1.2), что, в свою очередь, позволяет эффективно вычислять нижнюю оценку (2.6).

Нижнюю оценку (2.6) можно существенно уточнить при помощи перехода от отбрасывания ограничений целостности (2.2) к применению к ним релаксации Лагранжа (аналогично алгоритмам двойственной декомпозиции, см. раздел 1.3.2.2.2). Рассмотрим следующий лагранжиан:

$$L(\boldsymbol{y}, \boldsymbol{\lambda}) = \sum_{p \in \mathcal{P}} E_I^p(\boldsymbol{y}_p) + \sum_{j \in \mathcal{V}} \lambda_j \left(\sum_{p \in \mathcal{P}} y_{jp} - 1 \right). \tag{2.7}$$

Соотношение между минимумом максимумов (минимакс) и максимумом минимумов (максимин) позволяет получить следующую нижнюю оценку на точное решение задачи (2.1)-(2.3):

$$\min_{\boldsymbol{y} \in (2.2),(2.3)} E_I(\boldsymbol{y}) = \min_{\boldsymbol{y} \in (2.3)} \max_{\boldsymbol{\lambda}} L(\boldsymbol{y}, \boldsymbol{\lambda}) \geqslant \max_{\boldsymbol{\lambda}} \min_{\boldsymbol{y} \in (2.3)} L(\boldsymbol{y}, \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda}). \tag{2.8}$$

Здесь $D(\boldsymbol{\lambda})$ – это *двойственная функция*:

$$D(\boldsymbol{\lambda}) = \sum_{p \in \mathcal{P}} \min_{\boldsymbol{y}_p \in \{0,1\}^{|\mathcal{V}|}} \left(E_I^p(\boldsymbol{y}_p) + \sum_{j \in \mathcal{V}} \lambda_j y_{jp} \right) - \sum_{j \in \mathcal{V}} \lambda_j. \tag{2.9}$$

Минимизация функций

$$\Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}) = E_I^p(\boldsymbol{y}_p) + \sum_{j \in \mathcal{V}} \lambda_j y_{jp} \tag{2.10}$$

по переменным $\boldsymbol{y}_p = \{y_{ip}\}_{i \in \mathcal{V}}$ может быть проведена при помощи алгоритмов поиска минимального разреза графа, поскольку множители Лагранжа λ_i влияют только на унарные потенциалы $\lambda_j y_{jp}$, которые, в свою очередь, не влияют на свойство субмодулярности. Таким образом, двойственная функция $D(\boldsymbol{\lambda})$ может быть эффективно вычислена.

Утверждение 5. *Двойственная функция $D(\boldsymbol{\lambda})$ вогнута и кусочно-линейна. Субградиент функции $D(\boldsymbol{\lambda})$ может быть вычислен следующим образом:*

$$\partial D(\boldsymbol{\lambda}) = \sum_{p \in \mathcal{P}} y_{ip}^* - 1, \tag{2.11}$$

где $\{y_{ip}^*\}_{i \in \mathcal{V}} = \arg\min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda})$.

Доказательство. Заметим, что множества $\{0, 1\}^{|\mathcal{V}|}$, по которым проводится минимизация при вычислении двойственной функции $D(\boldsymbol{\lambda})$ (2.9), конечны. При фиксированных $\boldsymbol{y}_p$, $p \in \mathcal{P}$, лагранжиан $L(\boldsymbol{y}, \boldsymbol{\lambda})$ (2.7) линеен по переменным $\boldsymbol{\lambda}$. Отсюда следует, что функция $D(\boldsymbol{\lambda})$ имеет вид минимума из конечного числа линейных функций: $\min_i \boldsymbol{a}_i^\intercal \boldsymbol{\lambda}$. Функции такого вида кусочно-линейны и вогнуты. Субградиент равен $\boldsymbol{a}_j$, где $j = \arg\min_i \boldsymbol{a}_i^\intercal \boldsymbol{\lambda}$. □

Задача поиска наиболее точной нижней оценки вида (2.8) эквивалентна задаче максимизации функции $D(\boldsymbol{\lambda})$ по переменным $\boldsymbol{\lambda}$. Поскольку функция $D(\boldsymbol{\lambda})$ вогнута и позволяет эффективно вычислять субградиент, то для её максимизации можно применять методы негладкой оптимизации первого порядка. Конкретные методы оптимизации, использованные в данной работе, будут рассмотрены в главе 4.

Заметим, что размерность пространства, по которому требуется осуществить оптимизацию (количество множителей Лагранжа), составляет $|\mathcal{V}|$, что меньше чем в методе DD TRW [73] как минимум в $|\mathcal{P}|$ раз. Этот факт даёт основания надеяться, что алгоритм SMD будет сходиться быстрее, чем алгоритм DD TRW. Теоретическое сравнение нижних оценок методов SMD и DD TRW, а также их экспериментальное сравнение проведены в главах 3 и 5, соответственно.

2.2. Энергии с потенциалами высоких порядков

Рассмотрим задачу минимизации энергии, состоящей из унарных потенциалов и потенциалов произвольных порядков (1.1). Запишем энергию при помощи индикаторных переменных y_{ip}:

$$E_I(\boldsymbol{y}) = \sum_{i \in \mathcal{V}} \sum_{p \in \mathcal{P}} \theta_{ip} y_{ip} + \sum_{C \in \mathcal{C}} \sum_{\boldsymbol{d} \in \mathcal{X}_C} \theta_{C,\boldsymbol{d}} \prod_{\ell \in C} y_{\ell p_\ell}. \tag{2.12}$$

Минимизация функции (1.1) по многозначных дискретным переменным $\boldsymbol{x}$ эквивалентна минимизации энергии (2.12) по бинарным переменным $\boldsymbol{y}$ при ограничениях целостности (1.7).

Пусть все потенциалы высоких порядков разреженные (см. раздел 1.3.2.2.3). По определению разреженных потенциалов все коэффициенты $\theta_{C,\boldsymbol{d}}$ отрицательны. В этом случае при помощи тождества

$$\left(-\prod_{\ell\in C} y_\ell\right) = \min_{z\in\{0,1\}} \left((|C|-1)z - \sum_{\ell\in C} z y_\ell\right), \tag{2.13}$$

можно преобразовать функцию бинарных переменных E_I (2.12) с потенциалами высоких порядков к парно-сепарабельной функции E_I^* от расширенного набора переменных так, что минимум E_I^* по расширенному множеству переменных совпадает с минимумом E_I по исходному множеству переменных:

$$\min_{\boldsymbol{y}\in(1.4)} E_I(\boldsymbol{y}) = \min_{\boldsymbol{z},\boldsymbol{y}:\boldsymbol{y}\in(1.4)} E_I^*(\boldsymbol{y},\boldsymbol{z}).$$

Данный прием был предложен в работе [68] для потенциалов порядка 3, а в работе [38] был сформулирован в форме (2.13) и обобщен на случай потенциалов произвольных порядков.

Для произвольного потенциала высокого порядка заданного шаблонами θ_C (см. раздел 1.3.2.2.3) для того, чтобы применить преобразование (2.13), можно вычесть константу $\max_{\boldsymbol{f}\in\mathcal{D}_C}\theta_{C,\boldsymbol{f}}$ из всех значений потенциалов (в том числе при тех конфигурациях, где стоит значение по умолчанию), одновременно прибавив её в качестве константного слагаемого к энергии. После этого преобразование (2.13) можно применить, но, вообще говоря, придется ввести в энергию экспоненциально много дополнительных переменных z: по одной для каждого ненулевого значения. В случае же разреженных потенциалов количество добавляемых переменных равно сумме мощностей множеств $\mathcal{D}_C$.

Энергию $E_I^*(\boldsymbol{y},\boldsymbol{z})$ можно записать в виде

$$E_I^*(\boldsymbol{y},\boldsymbol{z}) = \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\theta_{ip}y_{ip} - \sum_{C\in\mathcal{C}}\sum_{\boldsymbol{d}\in\mathcal{D}_C}\theta_{C,\boldsymbol{d}}\left((|C|-1)z_{C,\boldsymbol{d}} - \sum_{\ell\in C} y_{\ell d_\ell} z_{C,\boldsymbol{d}}\right). \tag{2.14}$$

Функция $E_I^*(\boldsymbol{y},\boldsymbol{z})$ парно-сепарабельна и субмодулярна относительно переменных $\boldsymbol{y}$ и $\boldsymbol{z}$ (утв. 1), а значит, если нет никаких дополнительных ограничений, может быть эффективно минимизирована при помощи алгоритмов построения минимального разреза в графе (см. раздел 1.3.1.2).

Добавив ограничения целостности (1.4) к задаче минимизации функционала (2.14), применим к ним релаксацию Лагранжа. Аналогично случаю парно-сепарабельной энергии (см. раздел 2.1) запишем лагранжиан

$$L(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda}) = E_I^*(\boldsymbol{y},\boldsymbol{z}) + \sum_{i\in\mathcal{V}}\lambda_i\left(\sum_{p\in\mathcal{P}} y_{ip} - 1\right) \tag{2.15}$$

и двойственную функцию

$$D(\boldsymbol{\lambda}) = \min_{\boldsymbol{y},\boldsymbol{z}\in\{0,1\}} L(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda}). \tag{2.16}$$

Лагранжиан $L(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda})$ является субмодулярной парно-сепарабельной функцией относительно переменных $(\boldsymbol{y},\boldsymbol{z})$ для любых значений множителей Лагранжа $\boldsymbol{\lambda}$, что позволяет эффективно вычислить двойственную функцию D в произвольной точке $\boldsymbol{\lambda}$. Функция $D(\boldsymbol{\lambda})$ является нижней оценкой глобального минимума энергии (1.1). Аналогично утв. 5 функция $D(\boldsymbol{\lambda})$ переменных $\boldsymbol{\lambda}$ является вогнутой кусочно-линейной, что позволяет решать задачу максимизации $D(\boldsymbol{\lambda})$ при помощи алгоритмов выпуклой негладкой оптимизации. Субградиент функции $D(\boldsymbol{\lambda})$ может быть вычислен следующим образом:

$$\partial D(\boldsymbol{\lambda}) = \sum_{p\in\mathcal{P}} y^*_{ip} - 1, \tag{2.17}$$

где $(\{y^*_{ip}\}^{p\in\mathcal{P}}_{i\in\mathcal{V}}, \boldsymbol{z}^*) = \arg\min_{\boldsymbol{y}_p,\boldsymbol{z}\in\{0,1\}} L(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda})$. Возможность эффективно вычислить субградиент позволяет применять методы оптимизации первого порядка. Конкретные методы оптимизации, использованные в данной работе, будут рассмотрены в главе 4.

В случае парно-сепарабельной энергии с ассоциативными парными потенциалами SMR эквивалентно SMD. Если энергия парно-сепарабельна, но парные потенциалы не ассоциативны, то SMR можно существенно упростить. В этом случае вводить дополнительные переменные z при помощи (2.13) не требуется, поскольку лагранжиан (2.15)

$$L(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\lambda}) = \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}} \theta_{ip} y_{ip} + \sum_{\{i,j\}\in\mathcal{E}}\sum_{p,q\in\mathcal{P}} \left(\theta_{ij,pq} - \max_{k,\ell\in\mathcal{P}} \theta_{ij,k\ell}\right) y_{ip} y_{jq} + \sum_{i\in\mathcal{V}} \lambda_i \left(\sum_{p\in\mathcal{P}} y_{ip} - 1\right) + \theta_0$$

является парно-сепарабельным и субмодулярным уже относительно переменных $\boldsymbol{y}$, что позволяет эффективно вычислять двойственную функции (2.16). Здесь $\theta_0 = \sum_{\{i,j\}\in\mathcal{E}} \max_{k,\ell\in\mathcal{P}} \theta_{ij,k\ell}$.

Размерность пространства, в котором требуется осуществить оптимизацию, составляет $|\mathcal{V}|$, как и в методе SMD. Размерность пространства двойственных переменных, возникающего при использовании метода CWD [71], составляет не менее $|\mathcal{P}|(\sum_{C\in\mathcal{C}}|C| - |\mathcal{C}|)$. Метод PatB [71] позволяет сократить количество переменных путём усложнения подзадач, но, например, для гиперграфа, состоящего из 4-х связной решётки из парных потенциалов и дополнительных потенциалов высоких порядков, количество двойственных переменных составляет не менее $|\mathcal{P}|(|\mathcal{V}| + |\bigcup_{C\in\mathcal{C}:|C|>2} C|)$.

Робастные разреженные потенциалы. Выше в данном разделе подход SMR был применен для разреженных потенциалов высоких порядков. Данный подход можно обобщить и на случай *робастных разреженных потенциалов высоких порядков* при помощи методов, применённых в

работах [61, 104]. Потенциалы такого вида не только поощряют разметки тождественно соответствовать предопределённым конфигурациям $\boldsymbol{d} \in \mathcal{D}_C$, но и поощряют небольшие отклонения от них. Потенциал такого вида для гиперребра C можно выразить как через $|\mathcal{P}|$-значные переменные $\boldsymbol{x}$

$$\theta_C(\boldsymbol{x}_C) = \sum_{\boldsymbol{d} \in \mathcal{D}_C} \min \left(0,\ \theta_{C,\boldsymbol{d}} + \sum_{\ell \in C} w_\ell^C [x_\ell \neq d_\ell] \right),$$

так и через индикаторные переменные $\boldsymbol{y}$

$$\theta_C(\boldsymbol{y}_C) = \sum_{\boldsymbol{d} \in \mathcal{D}_C} \min \left(0,\ \theta_{C,\boldsymbol{d}} + \sum_{\ell \in C} w_\ell^C (1 - y_{\ell d_\ell}) \right).$$

Параметрами потенциалов гиперребра C являются константы $\theta_{C,\boldsymbol{d}} < 0$, $\boldsymbol{d} \in \mathcal{D}_C$ и $w_\ell^C > 0$, $\ell \in C$.

Последнее выражение может быть редуцировано к парно-сепарабельной функции при помощи добавления переменных-переключателей $z_{C,\boldsymbol{d}}$ (аналогично трансформации (2.14)):

$$\min_{z_{C,\boldsymbol{d}} \in \{0,1\}} \quad \theta_{C,\boldsymbol{d}} z_{C,\boldsymbol{d}} + \sum_{\ell \in C} w_\ell^C z_{C,\boldsymbol{d}} (1 - y_{\ell d_\ell}).$$

Данное выражение парно-сепарабельно и субмодулярно (т.к. $w_\ell^C \geqslant 0$) относительно переменных $\boldsymbol{y}$ и $\boldsymbol{z}$. Применение релаксации Лагранжа и максимизация двойственной функции в SMR с робастными разреженными потенциалами абсолютно аналогичны SMR с обычными разреженными потенциалами.

Для вывода SMR для робастных разреженных потенциалов использовалась формулировка из работы [104], а именно свёртка элементов различных выделенных конфигураций при помощи операции «сумма», а не «минимум». Данные две схемы эквивалентны, если все выделенные конфигурации достаточно далеко друг от друга (например, см. [61, ур. 17, 18]).

2.3. Несубмодулярный лагранжиан

В разделах 2.1 и 2.2 лагранжиан всегда был субмодулярен относительно бинарных переменных. Оказывается, что от этого требования можно отказаться, если использовать приближённый алгоритм минимизации энергии QPBO [66, 22] вместо точного алгоритма, основанного на построении минимально разреза графа.

Разберём данную ситуацию на примере задачи минимизации парно-сепарабельной энергии (2.1)-(2.3), где парные потенциалы θ_{ij} не являются ассоциативными (но, по-прежнему, $\theta_{ij,pq} = 0$ при $p \neq q$). В этом случае метод SMD (см. раздел 2.1) не применим. Для применения метода SMR (см. раздел 2.2) необходимо из всех значений потенциала вычесть его максимум: $\max_{p,q} \theta_{ij,pq}$. Данная операция может негативно сказаться на разреженности потенциала,

например, если он штрафует появление нескольких выделенных конфигураций. Если отбросить шаг вычитания максимума, то свойство разреженности сохраняется, но выражение (2.1) может быть несубмодулярно, а значит, минимизация его по бинарным переменным y_{ip} может быть NP-трудной задачей. Вместо того, чтобы выполнять минимизацию псевдо-булевой функции (2.1), можно вычислить её стандартную релаксацию при помощи алгоритма QPBO и далее применять релаксацию Лагранжа потенциалов целостности (2.2). Назовём данный подход *несубмодулярной релаксацией*, или, сокращённо, NSMR.

Приведем формальное описание NSMR. Рассмотрим следующий лагранжиан:

$$L(\boldsymbol{y}, \boldsymbol{\lambda}) = \sum_{i\in\mathcal{V}}\sum_{i\in\mathcal{P}} \theta_{ip} y_{ip} + \sum_{\{i,j\}\in\mathcal{E}} \sum_{(p,q)\in\mathcal{D}_{\{i,j\}}} \theta_{ij,pq} y_{ip} y_{jq} + \sum_{i\in\mathcal{V}} \lambda_i \left(\sum_{p\in\mathcal{P}} y_{ip} - 1 \right) \tag{2.18}$$

Лагранжиан $L(\boldsymbol{y}, \boldsymbol{\lambda})$ является парно-сепарабельной псевдо-булевой функцией переменных $\boldsymbol{y}$, а значит, его стандартная релаксация может быть вычислена при помощи алгоритма QPBO. Данный метод предоставляет значение нижней оценки $D_{QPBO}(\boldsymbol{\lambda})$ глобального минимума, а также *частичную разметку* переменных $\boldsymbol{y}$: каждой переменной y_{ip} ставится в соответствие значение из множества $\{0, 1, \emptyset\}$. Нижняя оценка, предоставляемая методом QPBO, по значению равна стандартной релаксации [132]:

$$D_{QPBO}(\boldsymbol{\lambda}) = \min_{\boldsymbol{y}\in \mathrm{Local}(\mathcal{G}')} L(\boldsymbol{y}, \boldsymbol{\lambda}), \tag{2.19}$$

а значит, функция $D_{QPBO}(\boldsymbol{\lambda})$ вогнута и кусочно-линейна, как функция переменных $\boldsymbol{\lambda}$. Здесь вершины графа $\mathcal{G}'$ соответствуют переменным y_{ip} лагранжиана (2.18), а ребра графа $\mathcal{E}'$ соответствуют слагаемым $\theta_{ij,pq} y_{ip} y_{jq}$ лагранжиана (2.18).

Для того чтобы вычислить субградиент функции $D_{QPBO}(\boldsymbol{\lambda})$, воспользуемся тем, что стандартная релаксация парно-сепарабельной псевдо-булевой функции является *полуцелой* [22]: существует решение задачи линейного программирования, такое что все переменные принимают значения из множества $\{0, 1, 0.5\}$. В частности, если QPBO присваивает значения 0, 1, $\emptyset$ переменной y_{ip}, то соответствующие переменные стандартной релаксации равны 0, 1, 0.5, соответственно. Парные потенциалы стандартной релаксации $y_{ij,pq}$ могут быть найдены по следующему правилу (док-во следует из леммы 1):

$$y_{ij,pq} = \begin{cases} \min(y_{ip}, y_{jp}), & \theta_{ij,pq} \leqslant 0, \\ \max(0, y_{ip} + y_{jp} - 1), & \text{иначе.} \end{cases} \tag{2.20}$$

Нижняя оценка $D_{QPBO}(\boldsymbol{\lambda})$, получаемая методом NSMR, является нижней оценкой нижней оценки, получаемой методом SMD, а значит, может привести к увеличению зазора между прямой и двойственной задачами. Данные свойства теоретически исследованы в разделе 3.4.

2.4. Линейные глобальные ограничения

Важным частным случаем глобальных потенциалов энергии являются линейные ограничения на индикаторные переменные. Можно перечислить несколько видов таких ограничений:

$$\sum_{j\in\mathcal{V}} y_{jp} = c, \tag{2.21}$$

$$\sum_{j\in\mathcal{V}} y_{jp} \in [c_1, c_2]. \tag{2.22}$$

Ограничения (2.21) являются жёсткими ограничениями на количество переменных, принимающих значение $p \in \mathcal{P}$. Ограничения (2.22) – интервальными ограничениями.

Пусть каждой вершине $j \in \mathcal{V}$ соответствует некоторая наблюдаемая скалярная или векторная величина I_j. Например, в случае если множество вершин $\mathcal{V}$ соответствует пикселям изображения, то величины I_j могут соответствовать интенсивностям или цветам (например, в пространстве RGB) пикселей изображения. Используя I_j в качестве весов, можно записать, например, следующие линейные ограничения на индикаторные переменные $\boldsymbol{y}$:

$$\sum_{j\in\mathcal{V}} y_{jp} I_j = \boldsymbol{\mu} \sum_{i\in\mathcal{V}} y_{ip}, \tag{2.23}$$

$$\sum_{j\in\mathcal{V}} y_{jp} (I_j - \boldsymbol{\mu})(I_j - \boldsymbol{\mu})^T = \sum_{j\in\mathcal{V}} y_{ip} \Sigma, \tag{2.24}$$

$$\sum_{j\in\mathcal{V}} y_{jp} I_j = \sum_{j\in\mathcal{V}} y_{jq} I_j, \tag{2.25}$$

Ограничения (2.23) и (2.24) задают равенство средних и ковариаций величин I_j величинам $\boldsymbol{\mu}$ и Σ, соответственно. Ограничение (2.25) задаёт равенство потоков через вершины, принимающие значения p и q.

Для некоторых частных случаев линейных ограничений существуют специализированные методы. Например, работы [91, 76] исследуют ограничения видов (2.21) и (2.22).

Подход SMR (а также SMD и NSMR) позволяет учитывать ограничения общего вида:

$$\sum_{i\in\mathcal{V}} \sum_{p\in\mathcal{P}} w_{ip}^m y_{ip} = c^m, \quad m = 1, \ldots, M \tag{2.26}$$

$$\sum_{i\in\mathcal{V}} \sum_{p\in\mathcal{P}} v_{ip}^k y_{ip} \leqslant d^k, \quad k = 1, \ldots, K. \tag{2.27}$$

Введём дополнительные множители Лагранжа $\boldsymbol{\xi} = \{\xi_m\}_{m=1,\ldots,M}$, $\boldsymbol{\pi} = \{\pi_m\}_{m=1,\ldots,M}$ по одному на каждое из ограничений (2.26), (2.27) и получим лагранжиан:

$$\begin{aligned} L(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi}) =& E_I(\boldsymbol{y}) + \sum_{i \in \mathcal{V}} \lambda_i \left(\sum_{p \in \mathcal{P}} y_{ip} - 1 \right) + \\ & \sum_{m=1}^{M} \xi_m \left(\sum_{j \in \mathcal{V}} \sum_{p \in \mathcal{P}} w_{ip}^m y_{ip} - c^m \right) + \\ & \sum_{k=1}^{K} \pi_k \left(\sum_{j \in \mathcal{V}} \sum_{p \in \mathcal{P}} v_{ip}^k y_{ip} - d^k \right), \end{aligned}$$

где множители Лагранжа, соответствующие ограничениям вида неравенство $\boldsymbol{\pi}$, должны быть неотрицательны. Лагранжиан субмодулярен относительно индикаторных переменных $\boldsymbol{y}$, и поэтому двойственная функция

$$D(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi}) = \min_{\boldsymbol{y}} L(\boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi}) \tag{2.28}$$

может быть эффективно вычислена.

Аналогично случаям, рассмотренным ранее, задача поиска наиболее точной нижней оценки соответствует задаче максимизации

$$\max_{\boldsymbol{\lambda},\, \boldsymbol{\xi},\, \boldsymbol{\pi} \geqslant 0} D(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi}).$$

Двойственная функция $D(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi})$ является кусочно-линейной и вогнутой по всем двойственным переменным, а значит, может быть максимизирована, например, методами выпуклой оптимизации.

3. Точность нижних оценок

В главе 2 настоящей работы получены методы для задачи минимизации энергии (1.1) общего вида, а также для нескольких частных случаев. Все описанные методы основаны на максимизации нижней оценки глобального минимума, полученной при помощи релаксации Лагранжа ограничений целостности (1.7). Качество этих методов существенно зависит от того, насколько точной является полученная оценка. В данной главе данный вопрос изучается теоретически: для каждого метода формулируется линейная релаксация, значению решения которой равен максимум его нижней оценки, а также проводится сравнение со стандартной релаксацией.

Сначала (раздел 3.1) приведены формулировки и доказательства нескольких вспомогательных лемм, облегчающих работу со стандартной линейной релаксацией функций бинарных переменных. Далее, раздел 3.2 посвящён парно-сепарабельным ассоциативным энергиям (метод SMD), раздел 3.3 – энергиям с потенциалами высоких порядков (метод SMR), раздел 3.4 – произвольным парно-сепарабельным энергиям (метод NSMR). В разделе 3.5 рассматривается случай глобальных линейных ограничений.

3.1. Вспомогательные леммы

Лемма 1. *Рассмотрим задачу минимизации парно-сепарабельной энергии*

$$E(\boldsymbol{y}) = \sum_{i \in \mathcal{A}} a_i y_i + \sum_{\{i,j\} \in \mathcal{B}} b_{ij} y_i y_j,$$

где переменные $\boldsymbol{y} = \{y_i\}_{i \in \mathcal{A}}$ *бинарны. Здесь* $\mathcal{A}$ *и* $\mathcal{B}$ *– множества унарных и парных потенциалов, соответственно. Тогда стандартная линейная релаксация данной энергии эквивалентна следующей задаче линейного программирования:*

$$\min_{y_i, y_{ij}} \quad \sum_{i \in \mathcal{A}} a_i y_i + \sum_{\{i,j\} \in \mathcal{B}} b_{ij} y_{ij}, \tag{3.1}$$

$$\text{s.t.} \quad y_\ell, y_{ij} \in [0,1], \quad \forall \ell \in \mathcal{A},\ \forall \{i,j\} \in \mathcal{B}, \tag{3.2}$$

$$y_{ij} \leqslant y_i,\ y_{ij} \leqslant y_j, \quad \forall \{i,j\} \in \mathcal{B}, \tag{3.3}$$

$$y_{ij} \geqslant y_i + y_j - 1, \quad \forall \{i,j\} \in \mathcal{B}. \tag{3.4}$$

Доказательство. Рассмотрим стандартную линейную релаксацию для рассматриваемой задачи:

$$\min_{\boldsymbol{z}} \quad \sum_{i\in\mathcal{A}} a_i z_{i1} + \sum_{\{i,j\}\in\mathcal{B}} b_{ij} z_{ij,11}, \tag{3.5}$$

$$\text{s.t.} \quad z_{i1}, z_{i0}, z_{ij,11}, z_{ij,01}, z_{ij,10}, z_{ij,00} \in [0,1], \tag{3.6}$$

$$z_{i1} + z_{i0} = 1, \tag{3.7}$$

$$z_{ij,10} + z_{ij,11} = z_{i1}, \quad z_{ij,01} + z_{ij,00} = z_{i0},$$

$$z_{ij,01} + z_{ij,11} = z_{j1}, \quad z_{ij,00} + z_{ij,10} = z_{j0} \tag{3.8}$$

Покажем, что по решению задачи (3.5)-(3.8) можно построить допустимую точку задачи (3.1)-(3.4) с таким же значением целевой функции, откуда будет следовать, что значение минимума задачи (3.1)-(3.4) не больше значения минимума задачи (3.5)-(3.8). Положим значения переменных следующим образом: $y_i = z_{i1}$, $y_{ij} = z_{ij,11}$. Допустимость данной точки следует из неотрицательности $\boldsymbol{z}$ и ограничений (3.8). Значения целевых функций обеих задач при таком присваивании совпадают.

Обратное соотношение между значениями минимумов задач (3.5)-(3.8) и (3.1)-(3.4) показывается аналогично. Действительно, по решению (3.1)-(3.4) можно положить $z_{j1} = y_j$, $z_{j0} = 1 - z_{j1}$, $z_{ij,11} = y_{ij}$, $z_{ij,01} = z_{j1} - z_{ij,11}$, $z_{ij,10} = y_{i1} - z_{ij,11}$, $z_{ij,00} = z_{ij,11} + 1 - y_{i1} - y_{j1}$. Легко убедиться, что такое присваивание переменных допустимо и приводит к такому же значению целевой функции.

Таким образом, показано, что значения минимумов двух задач совпадают, и по решению любой из них можно построить допустимую точку другой с таким же значением целевой функции, что и является эквивалентностью. □

Лемма 2. *Рассмотрим задачу минимизации парно-сепарабельной энергии*

$$E(\boldsymbol{y}) = \sum_{i\in\mathcal{A}} a_i y_i + \sum_{\{i,j\}\in\mathcal{B}} b_{ij} y_i y_j,$$

где переменные $\boldsymbol{y} = \{y_i\}_{i\in\mathcal{A}}$ бинарны. Здесь $\mathcal{A}$ и $\mathcal{B}$ – множества унарных и парных потенциалов, соответственно. Пусть все коэффициенты $b_{ij} \leqslant 0$. Тогда стандартная линейная релаксация данной энергии точна, и может быть записана как задача линейного программирования (3.1)-(3.3).

Доказательство. По утв. 1 функция $E(\boldsymbol{y})$ субмодулярна, а значит, согласно утв. 3, стандартная линейная релаксация данной задачи точна.

По лемме 1 стандартная линейная релаксация рассматриваемой задачи может быть записана в виде задачи линейного программирования (3.1)-(3.4). Покажем, что в данном случае

ограничение (3.4) можно опустить. Поскольку все $b_{ij} \leqslant 0$, то переменным y_{ij}, согласно целевой функции, выгоднее принимать наибольшее из допустимых значений, а именно $\min(y_i, y_j)$. Неравенство $\min(y_i, y_j) \geqslant y_i + y_j - 1$ выполнено для любых значений y_i и y_j, а значит ограничение (3.4) избыточно и может быть опущено. □

Лемма 3. *Рассмотрим два вектора $\boldsymbol{y}^1, \boldsymbol{y}^2 \in [0,1]^K$, $K \in \mathbb{N}$, таких что $\sum_{p=1}^K y_p^1 = \sum_{p=1}^K y_p^2 = 1$. Тогда существует набор значений $\{z_{pq}\}_{p,q=1}^K$, таких что*

$$
\begin{aligned}
& z_{pq} \geqslant 0, \quad \forall p, q \in \{1, \ldots, K\}, \\
& z_{pp} = \min(y_p^1, y_p^2), \quad \forall p, q \in \{1, \ldots, K\}, \\
& \sum_{q=1}^K z_{pq} = y_p^1, \quad \forall p \in \{1, \ldots, K\}, \\
& \sum_{p=1}^K z_{pq} = y_q^2, \quad \forall q \in \{1, \ldots, K\}.
\end{aligned}
$$

Доказательство. Положим $z_{pp}^* = \min(y_p^1, y_p^2)$. Рассмотрим транспортную задачу

$$
\begin{aligned}
\min_z \quad & \sum_{p \in \mathcal{P}} \sum_{q \in \mathcal{P}} z_{pq}, \\
\text{s.t.} \quad & \sum_{q \in \mathcal{P}} z_{pq} = a_p, \quad \forall p \in \mathcal{P}; \\
& \sum_{p \in \mathcal{P}} z_{pq} = b_q, \quad \forall q \in \mathcal{P}; \\
& z_{pq} \geqslant 0, \quad \forall p, q \in \mathcal{P},
\end{aligned}
$$

где $a_p = y_p^1 - z_{pp}^*$, $b_p = y_p^2 - z_{pp}^*$. Решение данной транспортной задачи $\{z_{pq}^0\}$ существует, поскольку $a_p \geqslant 0$, $b_p \geqslant 0$, и ограничения сбалансированы:

$$
\sum_{p \in \mathcal{P}} a_p = \sum_{p \in \mathcal{P}} (y_p^1 - z_{pp}^*) = 1 - \sum_{p \in \mathcal{P}} z_{pp}^* = \sum_{p \in \mathcal{P}} (y_p^2 - z_{pp}^*) = \sum_{p \in \mathcal{P}} b_p.
$$

При этом выполнено $z_{pp}^0 = 0$. Положим $z_{pq}^* = z_{pq}^0$, $p \neq q$. Набор значений $\{z_{pq}^*\}_{p,q=1}^K$ является искомым. □

Лемма 4. *Пусть $y_1, \ldots, y_K$ – произвольные действительные числа, принадлежащие отрезку $[0,1]$, $K > 1$. Рассмотрим задачу линейного программирования:*

$$
\min_{z, z_1, \ldots, z_K \in [0,1]} \quad z(K-1) - \sum_{k=1}^K z_k \tag{3.9}
$$

$$
\text{s.t.} \quad z_k \leqslant z, \; z_k \leqslant y_k. \tag{3.10}
$$

Минимальное значение целевой функции на допустимом множестве достигается в точке $z = z_1 = \cdots = z_K = \min_{k=1,\ldots,K} y_k$.

Доказательство. Не ограничивая общности, предположим, что $y_1 \leqslant \ldots \leqslant y_K$. Обозначим значения переменных, являющихся решением оптимизационной задачи (3.9)-(3.10), через $z^*, z_1^*, \ldots, z_K^*$, а значение минимума – через f^*. Переменная z входит в линейную целевую функцию (3.9) с положительным коэффициентом и ограничена неравенствами (3.10) снизу, что означает, что в точке оптимума она принимает наименьшее допустимое значение: $z^* = \max_{k=1,\ldots,K} z_k^*$. Аналогично, переменные $z_1, \ldots, z_K$ принимают наибольшие допустимые значения: $z_k^* = \min(y_k, z^*)$.

Покажем, что $z^* \geqslant y_1$. Предположим, что это не так. Пусть $\varepsilon = y_1 - z^* > 0$. Заметим, что точка $z^* + \varepsilon$, $z_1^* + \varepsilon, \ldots, z_K^* + \varepsilon$ является допустимой, целевая функция (3.9) в ней принимает значения $f^* - \varepsilon$, что противоречит предположению минимальности f^*.

Покажем, что $z^* \leqslant y_2$. Предположим, что это не так. Возможны 2 случая:

1. $y_K < z^*$. Пусть $\varepsilon = z^* - y_K > 0$. Точка $z^* - \varepsilon$, $z_1^*, \ldots, z_K^*$ является допустимой, целевая функция (3.9) в ней принимает значения $f^* - \varepsilon$, что противоречит предположению минимальности f^*.
2. $y_\ell < z^* \leqslant y_{\ell+1}$, для некоторого $\ell \in \{2, \ldots, K\}$. Пусть $\varepsilon = z^* - y_\ell > 0$. Точка $z^* - \varepsilon$, $z_1^*, \ldots, z_\ell^*, z_{\ell+1}^* - \varepsilon, \ldots, z_K^* - \varepsilon$ является допустимой, целевая функция (3.9) в ней принимает значения $f^* - \varepsilon(\ell - 1)$, что противоречит предположению минимальности f^*.

Все точки вида $z^* \in [y_1, y_2]$, $z_1^* = y_1$, $z_2^* = \cdots = z_K^* = z^*$ являются допустимыми, и на них целевая функция (3.9) принимает одинаковые значения, равные минимальному. Одной из таких точек и является точка, упомянутая в условии леммы. □

Лемма 5. *Рассмотрим задачу выпуклого программирования*

$$\min_{\boldsymbol{y} \in \mathcal{Y}} \quad f(\boldsymbol{y}) \tag{3.11}$$

$$\text{s.t.} \quad A\boldsymbol{y} = 0, \tag{3.12}$$

$$B\boldsymbol{y} \leqslant 0, \tag{3.13}$$

где $\mathcal{Y} \subseteq \mathbb{R}^n$ – выпуклое множество, $A : \mathbb{R}^n \to \mathbb{R}^m$ и $B : \mathbb{R}^n \to \mathbb{R}^k$ - линейные операторы, а $f : \mathbb{R}^n \to \mathbb{R}$ – выпуклая функция. Пусть множество $\mathcal{Y}$ обладает непустой относительной внутренностью[1], и значение решение оптимизационной задачи (3.11)-(3.13) *конечно.*

[1] Относительной внутренностью множества $\mathcal{Y}$ называется внутренняя область множества $\mathcal{Y}$ относительно линейного многообразия, являющегося аффинной оболочкой множества $\mathcal{Y}$. Более формально, $\text{relint}(\mathcal{Y}) = \{\boldsymbol{y} \mid \boldsymbol{y} \in \mathcal{Y},\ \exists \varepsilon > 0 : N_\varepsilon(\boldsymbol{y}) \cap \text{aff}(\mathcal{Y}) \subseteq \mathcal{Y}\}$, где $N_\varepsilon(\boldsymbol{y})$ – шар радиуса ε с центром в точке $\boldsymbol{y}$, а $\text{aff}(\mathcal{Y})$ – аффинная оболочка множества $\mathcal{Y}$.

Тогда выполнены следующие равенства:

$$\max_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^m\\ \boldsymbol{\mu}\in\mathbb{R}^k,\ \boldsymbol{\mu}\geqslant 0}} \min_{\boldsymbol{y}\in\mathcal{Y}} L(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\mu}) = \min_{\boldsymbol{y}\in\mathcal{Y}} \max_{\substack{\boldsymbol{\lambda}\in\mathbb{R}^m\\ \boldsymbol{\mu}\in\mathbb{R}^k,\ \boldsymbol{\mu}\geqslant 0}} L(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\mu}) = \min_{\substack{\boldsymbol{y}\in\mathcal{Y}:\\ \boldsymbol{y}\in(3.12)-(3.13)}} f(\boldsymbol{y}),$$

где $L(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\mu})$ *– лагранжиан, равный*

$$L(\boldsymbol{y},\boldsymbol{\lambda},\boldsymbol{\mu})) = f(\boldsymbol{y}) + \boldsymbol{\lambda}^T A\boldsymbol{y} + \boldsymbol{\mu}^T B\boldsymbol{y}.$$

Доказательство. Данное утверждение является прямым следствием теоремы о сильной двойственности в формулировке из книги Бертшекаша и др. [18, теорема 6.4.2]. □

3.2. Парно-сепарабельные ассоциативные энергии

В работах [97, 98] показано, что оптимальное значение нижней оценки, достигаемое методом SMD, равно глобальному минимуму стандартной линейной релаксации (теорема 1), а также релаксации Кляйнберга-Тардош (KT) [58], формулируемой для задачи uniform metric labeling (теорема 2).

Теорема 1. *Для парно-сепарабельной энергии с ассоциативными парными потенциалами* (2.4) *наилучшая нижняя оценка на глобальный минимум энергии, получаемая алгоритмом SMD, равна значению минимума стандартной линейной релаксации.*

Доказательство. Лагранжиан (2.7) является парно-сепарабельной функцией переменных $\boldsymbol{y}$. В слагаемых лагранжиана (2.5) все константы Cij, p не отрицательны, а значит применима лемма 2 для задачи минимизации лагранжиана по переменным y. Запишем вычисление двойственной функции $D(\boldsymbol{\lambda})$ при помощи леммы 2:

$$D(\boldsymbol{\lambda}) = \min_{\boldsymbol{y}} L(\boldsymbol{y},\boldsymbol{\lambda}) = \min_{\boldsymbol{y}} \quad \sum_{p\in\mathcal{P}} \left(\sum_{j\in\mathcal{V}} \theta_{jp} y_{jp} - \sum_{\{i,j\}\in\mathcal{E}} C_{ij,p} y_{ij,p} \right) + \sum_{j\in\mathcal{V}} \lambda_j \left(\sum_{p\in\mathcal{P}} y_{jp} - 1 \right), \quad (3.14)$$

$$\text{s.t.} \quad y_{\ell q}, y_{ij,p} \in [0,1], \quad \forall \ell \in \mathcal{V},\ \forall p,q \in \mathcal{P},\ \forall \{i,j\} \in \mathcal{E};$$

$$y_{ij,p} \leqslant y_{ip},\ y_{ij,p} \leqslant y_{jp}, \quad \forall \{i,j\} \in \mathcal{E},\ \forall p \in \mathcal{P}.$$

Применяя лемму 5, получим, что максимум двойственной функции $D(\boldsymbol{\lambda})$ по переменным $\boldsymbol{\lambda}$ равен значению решения следующей задачи линейного программирования:

$$\min_{\boldsymbol{y}} \quad \sum_{p\in\mathcal{P}}\left(\sum_{j\in\mathcal{V}}\theta_{jp}y_{jp} - \sum_{\{i,j\}\in\mathcal{E}}C_{ij,p}y_{ij,p}\right), \tag{3.15}$$

$$\text{s.t.} \quad y_{\ell q}, y_{ij,p} \in [0,1], \quad \forall \ell \in \mathcal{V},\ \forall p,q \in \mathcal{P},\ \forall \{i,j\} \in \mathcal{E}; \tag{3.16}$$

$$y_{ij,p} \leqslant y_{ip},\ y_{ij,p} \leqslant y_{jp}, \quad \forall \{i,j\} \in \mathcal{E},\ \forall p \in \mathcal{P}; \tag{3.17}$$

$$\sum_{p\in\mathcal{P}} y_{ip} = 1, \quad \forall i \in \mathcal{V}. \tag{3.18}$$

Покажем, что задача линейного программирования (3.15)-(3.18) эквивалентна стандартной линейной релаксации (опр. 5). Проведём рассуждения, аналогичные доказательству леммы 1.

Пусть y^*_{ip}, $y^*_{ij,p}$ – решение задачи (3.15)-(3.18). Построим набор переменных y^0_{ip}, $y^0_{ij,pq}$, являющийся допустимой точкой стандартной линейной релаксации, и такой что значение целевой функции (1.23) будет равно значению (3.15).

Рассмотрим ребро $\{i,j\} \in \mathcal{E}$. Так как в целевой функции (3.15) $C_{ij,p} \geqslant 0$, то $y^*_{ij,p} = \min(y^*_{ip}, y^*_{jp})$. При помощи леммы 3 построим набор $\{z_{pq}\}_{p,q=1}^{|\mathcal{P}|}$ на основе векторов $\{y^*_{ip}\}_{p\in\mathcal{P}}$ и $\{y^*_{jp}\}_{p\in\mathcal{P}}$. Положим в качестве значений переменных стандартной линейной релаксации значения переменых z_{pq} и y^*_{ip}: $y^0_{ip} := y^*_{ip}$, $y^0_{ij,pq} = z_{pq}$. Построенная таким образом точка y^0_{ip}, $y^0_{ij,pq}$ является допустимой точкой стандартной линейной релаксации (все переменный принадлежат отрезку $[0,1]$, выполнены ограничения (1.24), (1.25), и (1.26)). Значение целевой функции в этой точке равно значению функции (3.15) в точке y^*_{ip}, $y^*_{ij,p}$ по построению.

Пусть y^0_{ip}, $y^0_{ij,pq}$ – решение задачи стандартной линейной релаксации. Положим $y^*_{ip} := y^0_{ip}$ и $y^*_{ij,p} := y^0_{ij,pq}$. Набор переменных y^*_{ip}, $y^*_{ij,p}$ является допустимой точкой задачи (3.15)-(3.18), а также на нём значение целевой функции (3.15) равно значению целевой функции стандартной линейной релаксации. □

Теорема 2. *Для парно-сепарабельной энергии с ассоциативными парными потенциалами* (2.4) *стандартная линейная релаксация эквивалента релаксации Кляйнберга-Тардош:*

$$\begin{aligned}
\min_{z_{ip},z_{ij,p}} \quad & \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\hat{\theta}_{ip}z_{ip} + \frac{1}{2}\sum_{\{i,j\}\in\mathcal{E}}\sum_{p\in\mathcal{P}}C_{ij,p}z_{ij,p}, \\
\text{s.t.} \quad & \sum_{p\in\mathcal{P}} z_{ip} = 1, \quad \forall i \in \mathcal{V}, \\
& z_{ij,p} \geq z_{ip} - z_{jp}, \quad z_{ij,p} \geq z_{jp} - z_{ip}, \quad \forall \{i,j\} \in \mathcal{E}, \quad \forall p \in \mathcal{P} \\
& z_{ij,p} \geq 0, \quad z_{ip} \geq 0,
\end{aligned}$$

где $\hat{\theta}_{ip} = \theta_{ip} - \frac{1}{2}\sum_{j:\{i,j\}\in\mathcal{E}} C_{ij,p}$.[2]

Доказательство. Преобразуем целевую функцию (1.23) стандартной линейной релаксации:

$$\begin{aligned} E_L(\boldsymbol{y}_L) &= \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}} \theta_{ip} y_{ip} - \sum_{(i,j)\in\mathcal{E}}\sum_{p\in\mathcal{P}} C_{ij,p} y_{ij,pp} \\ &= \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}} (\theta_{ip} - \frac{1}{2}\sum_{j:\{i,j\}\in\mathcal{E}} C_{ij,p}) y_{ip} + \sum_{(i,j)\in\mathcal{E}}\sum_{\substack{p,q\in\mathcal{P}:\\ p\neq q}} \frac{C_{ij,p}+C_{ij,q}}{2} y_{ij,pq} \end{aligned}$$

Для выполнения данного преобразования использованы условия допустимости (1.25) и (1.26) точки $\boldsymbol{y}$.

Пусть даны два вектора $\boldsymbol{y}^1, \boldsymbol{y}^2 \in [0,1]^{|\mathcal{P}|}$, такие что $\sum_{p\in\mathcal{P}} y^1_p = \sum_{p\in\mathcal{P}} y^2_p = 1$, и неотрицательные константы C_p, $p \in \mathcal{P}$. Определим

$$f(\boldsymbol{y}^1, \boldsymbol{y}^2) = \min_{\boldsymbol{z}} \quad \sum_{p\in\mathcal{P}}\sum_{q\in\mathcal{P}} d_{pq} z_{pq}, \tag{3.19}$$

$$\text{s.t.} \quad \sum_{q\in\mathcal{P}} z_{pq} = y^1_p, \quad \forall p \in \mathcal{P}; \tag{3.20}$$

$$\sum_{p\in\mathcal{P}} z_{pq} = y^2_q, \quad \forall q \in \mathcal{P}; \tag{3.21}$$

$$z_{pq} \geqslant 0, \quad \forall p, q \in \mathcal{P}, \tag{3.22}$$

где $d_{pq} = 0$ при $p = q$, и $d_{pq} = \frac{C_p + C_q}{2}$ иначе.

Для доказательства теоремы достаточно показать, что $f(\boldsymbol{y}^1, \boldsymbol{y}^2) = \frac{1}{2}\sum_{p=1}^{P} C_p |y^1_p - y^2_p|$.

Пусть $\{z_{pq}\}_{p,q\in\mathcal{P}}$ – это набор, построенный по лемме 3. Покажем, что на этом векторе достигается минимум (3.19). Пусть $\{z'_{pq}\}_{p,q\in\mathcal{P}}$ – ещё один набор, удовлетворяющий ограничениям (3.20), (3.21), и (3.22). В этом случае выполнено $z'_{rr} \leqslant z_{rr} = \min(y^1_r, y^2_r)$, а значит

$$\begin{aligned} \sum_{p\in\mathcal{P}}\sum_{q\in\mathcal{P}} d_{pq} z'_{pq} &= \frac{1}{2}\left(\sum_{p\in\mathcal{P}} C_p y^1_p + \sum_{q\in\mathcal{P}} C_q y^2_q - 2\sum_{r\in\mathcal{P}} C_r z'_{rr}\right) \\ &\geqslant \frac{1}{2}\left(\sum_{p\in\mathcal{P}} C_p y^1_p + \sum_{q\in\mathcal{P}} C_q y^2_q - 2\sum_{r\in\mathcal{P}} C_r z_{rr}\right) = \sum_{p\in\mathcal{P}}\sum_{q\in\mathcal{P}} d_{pq} z_{pq}. \end{aligned}$$

[2] Первые формулировка и доказательство данной теоремы предложены В. Колмогоровым в работах [97, 98].

Цепочка следующих равенств завершает доказательство:

$$\begin{aligned}\sum_{p\in\mathcal{P}}\sum_{q\in\mathcal{P}} d_{pq}z_{pq} &= \frac{1}{2}\left(\sum_{p\in\mathcal{P}} C_p y_p^1 + \sum_{q\in\mathcal{P}} C_q y_q^2 - 2\sum_{r\in\mathcal{P}} C_r z_{rr}\right)\\ &= \frac{1}{2}\left(\sum_{r\in\mathcal{P}} C_r \min(y_r^1, y_r^2) + \sum_{r\in\mathcal{P}} C_r \max(y_r^1, y_r^2) - 2\sum_{r\in\mathcal{P}} C_r \min(y_r^1, y_r^2)\right)\\ &= \frac{1}{2}\left(\sum_{r\in\mathcal{P}} C_r \max(y_r^1, y_r^2) - \sum_{r\in\mathcal{P}} C_r \min(y_r^1, y_r^2)\right)\\ &= \frac{1}{2}\sum_{r\in\mathcal{P}} C_r |y_r^1 - y_r^2|.\end{aligned}$$

□

3.3. Энергии с потенциалами высоких порядков

В данном разделе для нижней оценки, получаемой методом SMR (см. раздел 2.2), формулируется эквивалентная задача линейного программирования (теорема 3); показывается, что в общем случае данная нижняя оценка, не лучше, чем нижняя оценка, получаемая методом CWD (см. раздел 2.2), а также, что в некотором частном случае разреженных потенциалов, нижние оценки, полученные методами SMR и CWD совпадают: теорема 4.

Теорема 3. *Для энергии вида* (2.12) *с разреженными потенциалами высоких порядков алгоритм SMR вычисляет следующую нижнюю оценку на глобальный минимум энергии:*

$$\min_{\boldsymbol{y}} \quad \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}} \theta_{ip} y_{ip} + \sum_{C\in\mathcal{C}}\sum_{\boldsymbol{d}\in\mathcal{D}_C} \theta_{C,\boldsymbol{d}} y_{C,\boldsymbol{d}} \tag{3.23}$$

$$\text{s.t.} \quad y_{ip}, y_{C,\boldsymbol{d}} \in [0,1], \quad \forall i\in\mathcal{V},\ \forall p\in\mathcal{P},\ \forall C\in\mathcal{C},\ \forall \boldsymbol{d}\in\mathcal{D}_C \tag{3.24}$$

$$y_{C,\boldsymbol{d}} \leqslant y_{\ell d_\ell}, \quad \forall C\in\mathcal{C},\ \forall \boldsymbol{d}\in\mathcal{D}_C,\ \forall \ell\in C, \tag{3.25}$$

$$\sum_{p\in\mathcal{P}} y_{ip} = 1, \quad \forall i\in\mathcal{V}. \tag{3.26}$$

Доказательство. Из леммы 4 следует, что задачу оптимизации (3.23)-(3.26) можно переписать в следующей эквивалентной форме:

$$\min_{\boldsymbol{y},\boldsymbol{z}} \quad \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\theta_{ip}y_{ip} - \sum_{C\in\mathcal{C}}\sum_{\boldsymbol{d}\in\mathcal{D}_C}\theta_C(\boldsymbol{d})\left((|C|-1)z_{C,\boldsymbol{d}} - \sum_{\ell\in C} z^{\ell}_{C,\boldsymbol{d}}\right) \tag{3.27}$$

$$\text{s.t.} \quad y_{ip}, z_{C,\boldsymbol{d}} \in [0,1], \quad \forall i\in\mathcal{V},\ \forall p\in\mathcal{P},\ \forall C\in\mathcal{C},\ \forall \boldsymbol{d}\in\mathcal{D}_C \tag{3.28}$$

$$z^{\ell}_{C,\boldsymbol{d}} \in [0,1], \quad \forall C\in\mathcal{C},\ \forall \boldsymbol{d}\in\mathcal{D}_C,\ \forall \ell\in C, \tag{3.29}$$

$$z^{\ell}_{C,\boldsymbol{d}} \leqslant z_{C,\boldsymbol{d}}, \quad \forall C\in\mathcal{C},\ \forall \boldsymbol{d}\in\mathcal{D}_C,\ \forall \ell\in C, \tag{3.30}$$

$$z^{\ell}_{C,\boldsymbol{d}} \leq y_{\ell d_\ell}, \quad \forall C\in\mathcal{C},\ \forall \boldsymbol{d}\in\mathcal{D}_C,\ \forall \ell\in C, \tag{3.31}$$

$$\sum_{p\in\mathcal{P}} y_{ip} = 1, \quad \forall i\in\mathcal{V}. \tag{3.32}$$

Обозначим целевую функцию (3.27) за $Q(\boldsymbol{y},\boldsymbol{z})$. Переменные $z_{C,\boldsymbol{d}}$ входят в линейную функцию $Q(\boldsymbol{y},\boldsymbol{z})$ только с положительными коэффициентами, а значит в точке оптимума принимают наименьшее значение из возможных: $z_{C,\boldsymbol{d}} = \max_{\ell\in C} z^{\ell}_{C,\boldsymbol{d}}$. Аналогично, $z^{\ell}_{C,\boldsymbol{d}} = \min(z_{C,\boldsymbol{d}}, y_{\ell d_\ell})$.

Пусть

$$R(\boldsymbol{\lambda}) = \min_{\boldsymbol{y},\boldsymbol{z}\in(3.28)-(3.31)} \left(Q(\boldsymbol{y},\boldsymbol{z}) + \sum_{i\in\mathcal{V}}\lambda_i\left(\sum_{p\in\mathcal{P}} y_{ip} - 1\right)\right).$$

Для разреженных потенциалов (см. раздел 1.3.2.2.3) коэффициенты неположительны $\theta_C(\boldsymbol{d}) \leqslant 0$, из чего следует, что задача (3.27), (3.28)-(3.31) эквивалента стандартной линейной релаксации парно-сепарабельной субмодулярной функции бинарных переменных (2.14) (лемма 2). Согласно утв. 3 в данном случае стандартная линейная релаксация точна, а значит $R(\boldsymbol{\lambda}) = D(\boldsymbol{\lambda})$ в произвольной точке $\boldsymbol{\lambda}$. Рассмотрим точку $\boldsymbol{\lambda}^* = \arg\max R(\boldsymbol{\lambda})$. В данной ситуации применима лемма 5, а значит

$$\begin{aligned} R(\boldsymbol{\lambda}^*) &= \max_{\boldsymbol{\lambda}} \min_{\boldsymbol{y},\boldsymbol{z}\in(3.28)-(3.31)} \left(Q(\boldsymbol{y},\boldsymbol{z}) + \sum_{i\in\mathcal{V}}\lambda_i\left(\sum_{p\in\mathcal{P}} y_{ip} - 1\right)\right) \\ &= \min_{\boldsymbol{y},\boldsymbol{z}\in(3.28)-(3.31)} \max_{\boldsymbol{\lambda}} \left(Q(\boldsymbol{y},\boldsymbol{z}) + \sum_{i\in\mathcal{V}}\lambda_i\left(\sum_{p\in\mathcal{P}} y_{ip} - 1\right)\right). \end{aligned}$$

Отсюда следует, что существуют значения переменных $\boldsymbol{y}^*$, удовлетворяющие ограничениям (3.32) и доставляющие минимум правой стороны равенства, что в свою очередь означает, что $\boldsymbol{y}^*$ является решением задачи (3.27)-(3.32). Первое равенство означает, что $R(\boldsymbol{\lambda}^*)$ равно значению решения этой задачи линейного программирования. Доказательство теоремы завершается равенством $R(\boldsymbol{\lambda}^*) = D(\boldsymbol{\lambda}^*)$. □

3.3.1. Перестановочные потенциалы Поттса

Линейная релаксация (3.23)-(3.26) не является наиболее точной известной линейной релаксацией задачи минимизации энергии с разреженными потенциалами высоких порядков. Комодакис и Параджиос [71] сформулировали более точную релаксацию, которая получается при помощи добавления к (3.23)-(3.26) условий согласованности переменных $y_{C,\boldsymbol{p}}$ и $y_{\ell q}$:

$$\sum_{\boldsymbol{p}\in\mathcal{P}^C:\, p_\ell=q} y_{C,\boldsymbol{p}} = y_{\ell q}, \quad \forall C \in \mathcal{C},\ \forall q \in \mathcal{P},\ \forall \ell \in C. \tag{3.33}$$

В общем случае релаксация Комодакиса и Параджиоса является более точной, чем релаксация, доказанная в теореме 3. Тем не менее, можно показать, что в некоторых частных случаях эти две релаксации эквиваленты (а значит метод SMR решает релаксацию Комодакиса и Параджиоса).

Определение 6. *Потенциал θ_C, $C \in \mathcal{C}$ называется перестановочным потенциалом Поттса (permuted $\mathcal{P}^n$-Potts), если он разреженный, и*

$$\forall \boldsymbol{d}', \boldsymbol{d}'' \in \mathcal{D}_C :\ \boldsymbol{d}' \neq \boldsymbol{d}'' \ \Rightarrow\ d'_i \neq d''_i,\ \forall i \in C.$$

Это определение означает, что каждая пара переменная-значение может принадлежать только одной выделенной конфигурации множества $\mathcal{D}_C$. Частными случаями перестановочных потенциалов Поттса являются потенциалы Поттса высокого порядка, введённые в работе [60], а также ассоциативные парные потенциалы [123].

Оказывается, что если все потенциалы высоких порядков являются перестановочными потенциалами Поттса, то релаксация (3.23)-(3.26) эквивалентна релаксации Комодакиса и Параджиоса.

Теорема 4. *Для энергии вида* (2.12) *с потенциалами высоких порядков, являющимися перестановочными потенциалами Поттса, алгоритм SMR вычисляет нижнюю оценку на глобальный минимум энергии, равную оптимальному значению релаксации Комодакиса и Параджиоса [71].*

Доказательство. Теорема 3 формулирует нижнюю оценку на глобальный минимум энергии, вычисляемую методом SMR, как задачу линейного программирования. Релаксация Комодакиса и Параджиоса отличается от сформулированной релаксации лишь наличием ограничений (3.33). Покажем, что в случае перестановочных потенциалов Поттса ограничения (3.33) являются избыточными.

Действительно, в случае разреженных потенциалов высокого порядка целевая функция содержит только слагаемые, соответствующие множествам $\mathcal{D}_C$, причём с отрицательными коэффициентами. Это означает, что переменные $y_{C,\boldsymbol{p}}$, $\boldsymbol{p} \in \mathcal{D}_C$ стремятся принять максимально

возможные значения, а значит ограничения (3.33) можно заменить на следующие:

$$\sum_{\boldsymbol{p}\in\mathcal{D}_C:\ p_\ell=q} y_{C,\boldsymbol{p}} \leqslant y_{\ell q},\quad \forall C\in\mathcal{C},\ \forall q\in\mathcal{P},\ \forall \ell\in C. \tag{3.34}$$

Если все потенциалы являются перестановочными потенциалами Поттса, то суммы, стоящие в левых частях неравенств (3.34), содержат не более одного слагаемого каждая, а значит ограничения (3.34) следуют из ограничений (3.25) и (3.24). □

Можно заметить, что теорема 1 является следствием теоремы 4, поскольку ассоциативные потенциалы второго порядка являются перестановочными потенциалами Поттса, а ограничения (3.33) переходят в условия согласованности (1.25), (1.26).

3.4. Произвольные парно-сепарабельные энергии

В данном разделе формулируется нижняя оценка, получаемая методом NSMR, изложенным в разделе 2.3.

Теорема 5. *Для парно-сепарабельной энергии алгоритм NSMR позволяет получить нижнюю оценку, равную значению решения следующей задачи линейного программирования:*

$$\min_{\boldsymbol{y}} \quad \sum_{i\in\mathcal{V}}\sum_{p\in\mathcal{P}}\theta_{ip}y_{ip} + \sum_{\{i,j\}\in\mathcal{E}}\ \sum_{(p,q)\in\mathcal{D}_{\{i,j\}}}\theta_{ij,pq}y_{ij,pq}, \tag{3.35}$$

$$\text{s.t. } y_{ip}, y_{ij,pq}\in[0,1], \tag{3.36}$$

$$y_{ij,pq}\leqslant y_{ip},\ y_{ij,pq}\leqslant y_{jq},\ \forall\{i,j\}\in\mathcal{E},\forall(p,q)\in\mathcal{D}_{\{i,j\}}, \tag{3.37}$$

$$y_{ij,pq}\geqslant y_{ip}+y_{jq}-1,\ \forall\{i,j\}\in\mathcal{E},\forall(p,q)\in\mathcal{D}_{\{i,j\}}, \tag{3.38}$$

$$\sum_{p\in\mathcal{P}} y_{ip}=1,\quad \forall i\in\mathcal{V}. \tag{3.39}$$

Доказательство. Известно, что алгоритм QPBO находит нижнюю оценку, равную значению минимума стандартной линейной релаксации. Согласно лемме 1 нижнюю оценку $D_{QPBO}(\boldsymbol{\lambda})$ (2.19) можно записать следующим образом:

$$\begin{aligned} D_{QPBO}(\boldsymbol{\lambda}) = \min_{\boldsymbol{y}} \quad & L(\boldsymbol{y},\boldsymbol{\lambda}),\\ \text{s.t.} \quad & y_{ip}\in[0,1],\quad \forall i\in\mathcal{V},\ \forall p\in\mathcal{P},\\ & y_{ij,pq}\in[0,1],\quad \forall\{i,j\}\in\mathcal{E},\ (p,q)\in\mathcal{D}_{\{i,j\}},\\ & y_{ij,pq}\leqslant y_{ip},\ y_{ij,pq}\leqslant y_{jq},\quad \forall\{i,j\}\in\mathcal{E},\ (p,q)\in\mathcal{D}_{\{i,j\}},\\ & y_{ij,pq}\geqslant y_{ip}+y_{jq}-1,\quad \forall\{i,j\}\in\mathcal{E},\ (p,q)\in\mathcal{D}_{\{i,j\}}, \end{aligned}$$

где $L(\boldsymbol{y}, \boldsymbol{\lambda})$ – лагранжиан (2.18).

Применяя лемму 5, получаем следующее равенство

$$\max_{\boldsymbol{\lambda}} \min_{\boldsymbol{y} \in (3.36)-(3.38)} L(\boldsymbol{y}, \boldsymbol{\lambda}) = \min_{\boldsymbol{y} \in (3.36)-(3.38)} \max_{\boldsymbol{\lambda}} L(\boldsymbol{y}, \boldsymbol{\lambda})$$
$$= \min_{\boldsymbol{y} \in (3.36)-(3.39)} \sum_{i \in \mathcal{V}} \sum_{p \in \mathcal{P}} \theta_{ip} y_{ip} + \sum_{\{i,j\} \in \mathcal{E}} \sum_{p,q \in \mathcal{P}} \theta_{ij,pq} y_{ij,pq},$$

что и требовалось доказать. □

3.5. Линейные глобальные ограничения

Нижние оценки, получаемые методами SMR, SMD, и NSMR с глобальными линейными ограничениями, равны решениям задач линейного программирования, получаемых при помощи соответствующих методов без ограничений, с добавлением соответствующих линейных равенств и неравенств. Ниже приведена соответствующая теорема для метода SMR.

Теорема 6. *Наилучшая нижняя оценка на значение решения задачи минимизации энергии* (1.5)-(1.7) *при ограничениях* (2.26) *и* (2.27)*, построенная методом SMR (максимум функции* $D(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi})$ (2.28) *при ограничениях* $\pi_k \geqslant 0$*), равна значению решения задачи линейного программирования* (3.23)-(3.26) *с добавлением линейных ограничений* (2.26) *и* (2.27)*.*

Доказательство данной теоремы аналогично доказательству теоремы 3. Ключевым используемым фактом является лемма 5.

Результаты для методов SMD и NSMR формулируются и доказываются аналогично.

Теорема 7. *Наилучшая нижняя оценка на значение решения задачи минимизации парносепарабельной энергии* (2.1)-(2.3) *с ассоциативными потенциалами* (2.4) *при ограничениях* (2.26) *и* (2.27)*, построенная методом SMD (максимум функции* $D(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi})$ (2.28) *при ограничениях* $\pi_k \geqslant 0$*), равна значению решения стандартной линейной релаксации (опр. 5) с добавлением линейных ограничений* (2.26) *и* (2.27)*.*

Теорема 8. *Наилучшая нижняя оценка на значение решения задачи минимизации парносепарабельной энергии* (2.1)-(2.3) *с произвольными потенциалами при ограничениях* (2.26) *и* (2.27)*, построенная методом NSMR (максимум функции* $D(\boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\pi})$ (2.28) *при ограничениях* $\pi_k \geqslant 0$*), равна значению решения задачи линейного программирования* (3.35)-(3.39) *с добавлением линейных ограничений* (2.26) *и* (2.27)*.*

4. Максимизация нижних оценок

В данной главе обсуждаются теоретические и практические аспекты применения алгоритма SMR. Раздел 4.1 посвящён теоретическому анализу свойств точки максимума нижней оценки SMR. Разделы 4.2 и 4.3 описывают два подхода к решению задачи максимизации двойственной функции, возникающей в алгоритме SMR, а раздел 4.4 описывает практические алгоритмы построения прямого решения по некоторому значению двойственных переменных.

4.1. Теоретические свойства точек максимума

В рамках данного раздела будем считать, что для задачи минимизации энергии $E(\boldsymbol{x})$ (1.1) построена двойственная функция $D(\boldsymbol{\lambda}) = \min_{\boldsymbol{y}} L(\boldsymbol{y}, \boldsymbol{\lambda})$, которая является нижней оценкой на глобальный минимум энергии $E(\boldsymbol{x})$ для любого значения переменных $\boldsymbol{\lambda}$. При этом лагранжиан $L(\boldsymbol{y}, \boldsymbol{\lambda})$ является линейной функцией относительно вещественных переменных $\boldsymbol{\lambda}$ и субмодулярной функцией относительно бинарных переменных $\boldsymbol{y}$. Функция $D(\boldsymbol{\lambda})$ является вогнутой кусочно-линейной и может быть эффективно вычислена. Заметим, что описанная ситуация наблюдается, как в методе SMD (см. раздел 2.1), так и в его обобщении – методе SMR (см. раздел 2.2).

Введём несколько дополнительных определений и обозначений. Обозначим точку максимума двойственной функции $D(\boldsymbol{\lambda})$ через $\boldsymbol{\lambda}^*$, значение переменных $\boldsymbol{y}$, при которых достигается такое значение лагранжиана, – через $\boldsymbol{y}^*$: $\boldsymbol{\lambda}^* = \arg\max D(\boldsymbol{\lambda})$, $\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} L(\boldsymbol{y}, \boldsymbol{\lambda}^*)$.

Для произвольной точки $\boldsymbol{\lambda}$ обозначим множество значений y_{ip}, при которых может достигаться минимум лагранжиана $L(\boldsymbol{y}, \boldsymbol{\lambda})$ по переменным $\boldsymbol{y}$, через $Z_{ip}(\boldsymbol{\lambda})$:

$$Z_{ip}(\boldsymbol{\lambda}) = \left\{ z \mid \exists \hat{\boldsymbol{y}} \in \operatorname{Arg\,min}_{\boldsymbol{y}} L(\boldsymbol{y}, \boldsymbol{\lambda}) \ : \ \hat{y}_{ip} = z \right\}, \quad i \in \mathcal{V},\ p \in \mathcal{P}.$$

4.1.1. Условия сильной и слабой согласованности

По аналогии с понятиями сильной и слабой согласованности деревьев (strong tree agreement, STA и weak tree agreement, WTA) для алгоритмов TRW [130, 65] определим поня-

тия сильной согласованности (strong agreement, SA) и слабой согласованности (weak agreement, WA).

Определение 7. *Значение двойственных переменных* $\boldsymbol{\lambda}$ *удовлетворяет условию* сильной согласованности *(SA), если для каждой из исходных переменных* $i \in \mathcal{V}$ *ровно для одной метки* $p \in \mathcal{P}$ *множество* $Z_{ip}(\boldsymbol{\lambda})$ *содержит элемент* 1*, а для всех остальных меток* $q \neq p$ *множества* $Z_{iq}(\boldsymbol{\lambda})$ *содержат только элемент* 0*:*

$$\forall i \in \mathcal{V}\ \exists! p \in \mathcal{P}:\ 1 \in Z_{ip}(\boldsymbol{\lambda}),\ \ \forall q \neq p:\ Z_{iq}(\boldsymbol{\lambda}) = \{0\}.$$

Определение 8. *Значение двойственных переменных* $\boldsymbol{\lambda}$ *удовлетворяет условию* слабой согласованности *(WA), если для каждой вершины* $i \in \mathcal{V}$ *выполнены два условия:*

1. $\exists p \in \mathcal{P}:\ 1 \in Z_{ip}(\boldsymbol{\lambda})$.
2. $\forall p \in \mathcal{P}:\ Z_{ip}(\boldsymbol{\lambda}) = \{1\} \Rightarrow \forall q \neq p,\ 0 \in Z_{jq}(\boldsymbol{\lambda})$.

Определение слабой согласованности фактически означает, что для выбранной точки $\boldsymbol{\lambda}$ существует значение бинарных переменных $\boldsymbol{y}$, согласованное со всеми множествами $Z_{ip}(\boldsymbol{\lambda})$ и ограничениями целостности (1.7).

Докажем несколько простых утверждений, поясняющих суть определений сильной и слабой согласованности.

Утверждение 6. *Если точка* $\hat{\boldsymbol{\lambda}}$ *удовлетворяет условию сильной согласованности, то значение* $D(\hat{\boldsymbol{\lambda}})$ *равно глобальному минимуму энергии* $E(\boldsymbol{x})$ *и конфигурация* $\hat{\boldsymbol{x}}$*, на которой достигается глобальный минимум, однозначно задаётся множествами* $Z_{ip}(\hat{\boldsymbol{\lambda}})$*:* $\hat{x}_i = k$*, где* $1 \in Z_{ik}(\hat{\boldsymbol{\lambda}})$.

Доказательство. Из определения сильной согласованности следует, что для точки $\hat{\boldsymbol{\lambda}}$ существует единственный бинарный вектор $\hat{\boldsymbol{y}}$, согласованный и с множествами $Z_{ip}(\hat{\boldsymbol{\lambda}})$: $\hat{y}_{ip} \in Z_{ip}(\hat{\boldsymbol{\lambda}})$, и с ограничениями целостности (1.7). Вектор $\hat{\boldsymbol{y}}$ однозначно задаёт вектор дискретных значений $\hat{\boldsymbol{x}} \in \mathcal{P}^{\mathcal{V}}$, который является глобальным минимумом энергии, поскольку значение $E(\hat{\boldsymbol{x}})$ совпадает со значением лагранжиана $L(\hat{\boldsymbol{y}}, \hat{\boldsymbol{\lambda}})$. □

Утверждение 7. *Если точка* $\hat{\boldsymbol{\lambda}}$ *удовлетворяет условию сильной согласованности, то она удовлетворяет и условию слабой согласованности.*

Доказательство очевидно из определений сильной и слабой согласованности. Обратное утверждение неверно.

Утверждение 8. *Точка* $\boldsymbol{\lambda}^*$ *максимума двойственной функции* $D(\boldsymbol{\lambda})$ *удовлетворяет условию слабой согласованности.*

Доказательство. Проведём доказательство от противного. Пусть точка $\boldsymbol{\lambda}^*$ не удовлетворяет условию слабой согласованности. Тогда для некоторой вершины $j \in \mathcal{V}$ либо существуют метки $p \neq q$, такие что $Z_{jp}(\boldsymbol{\lambda}^*) = Z_{jq}(\boldsymbol{\lambda}^*) = \{1\}$, либо $Z_{jp}(\boldsymbol{\lambda}^0) = \{0\}$ выполнено для всех меток $p \in \mathcal{P}$.

В первом случае рассмотрим точку $\hat{\boldsymbol{\lambda}}$, такую что $\hat{\lambda}_i = \lambda_i^*$ для всех $i \neq j$ и $\hat{\lambda}_j = \lambda_j^* + \varepsilon$, где $\varepsilon > 0$. При достаточно малом ε элемент 1 по-прежнему входит в оба множества $Z_{jp}(\hat{\boldsymbol{\lambda}})$ и $Z_{jq}(\hat{\boldsymbol{\lambda}})$, а элемент 0 входит во все остальные множества $Z_{j\ell}(\hat{\boldsymbol{\lambda}})$, $\ell \in \mathcal{P} \setminus \{p, q\}$. Отсюда следует, что $D(\hat{\boldsymbol{\lambda}}) \geqslant D(\boldsymbol{\lambda}^*) + \varepsilon > D(\boldsymbol{\lambda}^*)$, что противоречит предположению максимальности $D(\boldsymbol{\lambda}^*)$.

Во втором случае рассмотрим точку $\hat{\boldsymbol{\lambda}}$, такую что $\hat{\lambda}_i = \lambda_i^*$ для всех $i \neq j$ и $\hat{\lambda}_j = \lambda_j^* - \varepsilon$, где $\varepsilon > 0$. При достаточно малом ε элемент 0 входит во все множества $Z_{j\ell}(\hat{\boldsymbol{\lambda}})$, $\ell \in \mathcal{P}$, откуда следует, что $D(\hat{\boldsymbol{\lambda}}) \geqslant D(\boldsymbol{\lambda}^*) + \varepsilon > D(\boldsymbol{\lambda}^*)$, что противоречит предположению максимальности $D(\boldsymbol{\lambda}^*)$. □

Обратное утверждение также неверно.

4.1.2. Зазор между прямой и двойственной задачами

Одним из наиболее важных показателей, определяющих успешность применения методов, основанных на максимизации двойственной функции, основанной на релаксации Лагранжа (таких как SMR и DD TRW), является размер зазора между прямой и двойственной задачами: $\min_{\boldsymbol{x}} E(\boldsymbol{x}) - \max_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda})$. Из утв. 6 следует, что при наличия точки сильного согласования величина зазора равна 0. При наличии же точки слабого согласования сказать что-либо о величине зазора нельзя.

Функция $D(\boldsymbol{\lambda})$ является вогнутой и кусочно-линейной, каждый линейный сегмент этой функции задаётся некоторым значением дискретных переменных $\boldsymbol{y}$. При этом, если какой-то из сегментов, задающих грань функции $D(\boldsymbol{\lambda})$, одновременно является «горизонтальным», то функция на этом сегменте достигает своего максимума. Более того, в этом случае зазор равен 0. Следующее утверждение формализует эти интуитивные рассуждения.

Утверждение 9. *Глобальный минимум энергии $E(\boldsymbol{x})$ равен значению лагранжиана $L(\boldsymbol{y}^*, \boldsymbol{\lambda}^*)$ в точке максимина $(\boldsymbol{y}^*, \boldsymbol{\lambda}^*)$ ($L(\boldsymbol{y}^*, \boldsymbol{\lambda}^*) = \max_{\boldsymbol{\lambda}} \min_{\boldsymbol{y} \in (1.6)} L(\boldsymbol{y}, \boldsymbol{\lambda})$) тогда и только тогда, когда существует бинарный вектор $\hat{\boldsymbol{y}}$, задающий горизонтальную гиперплоскость ($L(\hat{\boldsymbol{y}}, \boldsymbol{\lambda})$ не зависит от $\boldsymbol{\lambda}$), такой что $L(\hat{\boldsymbol{y}}, \boldsymbol{\lambda}^*) = L(\boldsymbol{y}^*, \boldsymbol{\lambda}^*)$.*

Доказательство. Пусть глобальный минимум энергии $E(\boldsymbol{x})$ равен значению $L(\boldsymbol{y}^*, \boldsymbol{\lambda}^*)$. По точке $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} E(\boldsymbol{x})$ построим вектор $\hat{y}$: $\hat{y}_{ip} = [\hat{x}_i = p]$. Вектор $\hat{y}$ и является искомым, поскольку удовлетворяет ограничениям целостности (1.7), а значит, $E(\hat{\boldsymbol{x}}) = L(\hat{\boldsymbol{y}}, \boldsymbol{\lambda}) = L(\boldsymbol{y}^*, \boldsymbol{\lambda}^*)$.

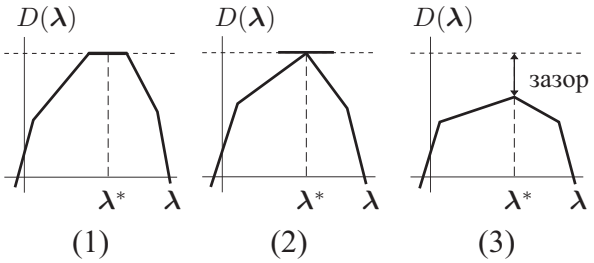


Рисунок 4.1.: Три ситуации, возможные в окрестности точки глобального максимума $\boldsymbol{\lambda}^*$ двойственной функции $D(\boldsymbol{\lambda})$. По осям абсцисс отложены значения переменных $\boldsymbol{\lambda}$ (здесь одномерные). По осям ординат отложены значения энергий и двойственных функций $D(\boldsymbol{\lambda})$. Значение глобального минимума энергии показано горизонтальной пунктирной линией. Значения двойственной функции в каждой точке $\boldsymbol{\lambda}$ показано жирной сплошной линией. Левый график (1) показывает ситуацию, когда удаётся найти и значение глобального минимума энергии, и разметку переменных, на которых оно достигается. Центральный график (2) показывает ситуацию, когда удалось найти значение глобального минимума энергии, но не конфигурацию переменных. Жирная горизонтальная линия соответствует разметке переменных $\hat{\boldsymbol{y}}$, которая существует, но, вообще говоря, неизвестна. Правый график (3) показывает ситуацию, когда не существует горизонтальной гиперплоскости, проходящей через точку оптимума. В этом случае существует ненулевой зазор.

Пусть теперь существует вектор $\hat{y}$, удовлетворяющий условиям задачи. Поскольку $L(\hat{\boldsymbol{y}}, \boldsymbol{\lambda}) = L(\hat{\boldsymbol{y}}, \boldsymbol{\lambda}^*)$, $\forall \boldsymbol{\lambda}$, то вектор $\hat{y}$ удовлетворяет ограничению (1.7), а значит, задаёт разметку $\hat{\boldsymbol{x}}$, значение энергии на которой равно максимину лагранжиана. □

Из утверждения 9 следует, что в окрестности точки $\boldsymbol{\lambda}^*$ максимума двойственной функции $D(\boldsymbol{\lambda})$ возможны 3 ситуации (см. рис. 4.1):

1. все множества $Z_{ip}(\boldsymbol{\lambda}^*)$ состоят из единственных элементов; бинарный вектор $\boldsymbol{y}^*$, удовлетворяющий ограничениям целостности (1.24) (а значит и оптимальная дискретная разметка $\boldsymbol{x}^*$) может быть в явном виде восстановлен; зазор между прямой и двойственной задачами равен 0;

2. некоторые, возможно все, множества $Z_{ip}(\boldsymbol{\lambda}^*)$ состоят из нескольких элементов; разметку $\boldsymbol{y}^*$, согласованную с (1.24) восстановить, вообще говоря, нетривиально; зазора между прямой и двойственной задачами по-прежнему нет;

3. некоторые множества $Z_{ip}(\boldsymbol{\lambda}^*)$ состоят из нескольких элементов; разметка $\boldsymbol{y}^*$, одновременно удовлетворяющая ограничениям (1.24) и доставляющая минимум лагранжиана не существует; зазор между прямой и двойственной задачами строго больше 0.

В ситуации 1 в точке $\boldsymbol{\lambda}^*$ выполнено условие сильной согласованности, а значит удаётся найти как значение глобального минимума энергии, так и конфигурацию переменных $\boldsymbol{x}$, на которой оно достигается. В ситуации 2 в точке $\boldsymbol{\lambda}^*$ выполнено условие слабой согласованности согласованности. При этом зазор между прямой и двойственной задачами равен 0, а значит оптимальная согласованная разметка $\hat{\boldsymbol{y}}$ существует, хоть может быть не известна. В ситуации 3 в точке $\boldsymbol{\lambda}^*$ также выполнено условие слабой согласованности, но существует ненулевой зазор между прямой и двойственной задачами. Бинарного вектора $\hat{\boldsymbol{y}}$, возникающего в утверждении 9 не существует.

В ситуации 3 алгоритм строящий оптимальные конфигурации $\boldsymbol{x}$ построить не удаётся. В ситуации 2 в некоторых частных случаях можно построить алгоритмы, строящие разметку $\boldsymbol{x}^*$, доставляющую глобальный минимум энергии (1.1). Например, таким частным случаем является ситуация, когда для каждой компоненты связности (относительно множества гиперрёбер $\mathcal{C}$) множества вершин, для которых есть неоднозначности в разметке, существует одна метка, которой можно покрыть всю компоненту. Более подробный анализ этой ситуации приведен в разделе 4.4.3.

4.2. Методы оптимизации для решения двойственной задачи

В данном разделе рассматриваются несколько методов оптимизации, которые применялись для решения задач максимизации различных нижних оценок на глобальный минимум энергии. Обсуждается возможность использования этих методов для поиска наилучшей нижней оценки из семейства оценок SMR (2.16). Ниже приведен список методов, ранее использованных для оптимизации функционалов, соответствующих поиску наилучших нижних оценок на решение задач минимизации энергий (в основном, DD TRW):

1. методы субградиентного подъёма [9, 73];
2. методы пучков субградиентов (bundle methods) [52];
3. алгоритм BFGS для негладких функций [90];
4. стохастический субградиентный метод [73].
5. проксимальные методы [139, 102];
6. методы, основанные на сглаживании функционала [106, 139];

Алгоритмы групп 1, 2, и 3 возможно применить для максимизации двойственной функции SMR, поскольку они требуют возможности вычислять только значение функции $D(\boldsymbol{\lambda})$ и один её субградиент $\partial D(\boldsymbol{\lambda})$ для произвольной точки $\boldsymbol{\lambda}$. Данные методы обладают большим числом деталей реализации и параметров, существенно влияющих как на качество, так и на скорость работы. Рассмотрим детали реализации этих методов.

Метод субградиентного подъёма. Субградиентные методы на каждой итерации делают шаг в направлении субградиента, вычисленного в текущей точке:

$$\boldsymbol{\lambda}^{n+1} = \boldsymbol{\lambda}^n + \alpha^n \partial D(\boldsymbol{\lambda}^n), \tag{4.1}$$

где числа α^n называются длиной шага, способ их выбора является важным параметром метода.

Можно показать, что метод субградиентного подъёма сходится к оптимуму функции, если все α^n положительны, ряд α^n расходится, но ряд квадратов α^n сходится [1, стр. 259] (обзор разных теорем о сходимости приведен, например, в работе [13]):

$$\alpha^n > 0, \quad \sum_{n=1}^{\infty} \alpha^n = \infty, \quad \sum_{n=1}^{\infty} (\alpha^n)^2 < \infty. \tag{4.2}$$

В работах [73] и [52] используются следующие варианты выбора последовательностей длин шагов:

$$\alpha^n = \frac{\gamma}{1 + \beta n} \tag{4.3}$$

$$\alpha^n = \frac{\gamma}{1 + \beta n} \frac{1}{\|\partial D(\boldsymbol{\lambda}^n)\|_2}, \tag{4.4}$$

где γ и β – параметры методов.

Также в работах [73] и [52] предлагается следующая адаптивная схема выбора длины шага α^n, которая не обладает гарантиями сходимости, но, тем не менее, часто хорошо работает на практике:

$$\alpha^n = \gamma \frac{A^n - D(\boldsymbol{\lambda}^n)}{\|\partial D(\boldsymbol{\lambda^n})\|_2^2}, \tag{4.5}$$

где A^n является текущей оценкой на оптимальное значение функционала (наименьшее значение энергии, найденное к текущей итерации), а γ – параметр метода.

Методы пучков субградиентов. Каппес и др. [52] применяют методы пучков субградиентов (bundle methods) [57, 84] для максимизации нижней оценки. Неформально методы данного типа можно описать как комбинацию приближения подграфика функции $D(\boldsymbol{\lambda})$ (выпуклого множества) при помощи кусочно-линейной поверхности (как в методе отсекающих плоскостей) и

проксимального оператора для предотвращения слишком больших шагов. Далее приведём формальное описание методов пучков субградиентов.

Пучок (bundle) $\mathcal{B}$ – это набор троек, состоящих из точек $\boldsymbol{\lambda}'$, значений оптимизируемой функции $D(\boldsymbol{\lambda}')$ в этих точках, и субградиентов в них $\partial D(\boldsymbol{\lambda}')$. На каждом шаге методы пучков решают следующую оптимизационную задачу:

$$\boldsymbol{\lambda}^{n+1} = \arg\max_{\boldsymbol{\lambda}} \left(\hat{D}(\boldsymbol{\lambda}) - \frac{w^n}{2} \|\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}\|_2^2 \right), \tag{4.6}$$

где $\hat{D}(\boldsymbol{\lambda})$ – верхняя оценка двойственной функции $D(\boldsymbol{\lambda})$:

$$\hat{D}(\boldsymbol{\lambda}) = \min_{(\boldsymbol{\lambda}', D(\boldsymbol{\lambda}'), \partial D(\boldsymbol{\lambda}')) \in \mathcal{B}} \{ D(\boldsymbol{\lambda}') + \langle \partial D(\boldsymbol{\lambda}'), \boldsymbol{\lambda} - \boldsymbol{\lambda}' \rangle \},$$

а $\bar{\boldsymbol{\lambda}}$ – текущее приближение оптимума.

Оптимизационную задачу (4.6) можно переформулировать как задачу квадратичного программирования небольшого размера (количество переменных и ограничений равно количеству элементов в пучке) и решить, например, при помощи методов внутренней точки (см. [52]). Если решение данной задачи не приводит к существенному увеличению оптимизируемой функции $D(\boldsymbol{\lambda})$, то происходит малый шаг (null step): в пучок $\mathcal{B}$ добавляется тройка $(\boldsymbol{\lambda}^{n+1}, D(\boldsymbol{\lambda}^{n+1}), \partial D(\boldsymbol{\lambda}^{n+1}))$. В противном случае происходит большой шаг (serious step): помимо обновления пучка аналогично малому шагу, происходит перенос оценки решения $\bar{\boldsymbol{\lambda}}$ в текущую точку $\boldsymbol{\lambda}^{n+1}$. Фактически, малые шаги соответствуют уточнению локального приближения функции, а большие шаги переносу центра в новую точку. В работах [84, 52] для выбора между большим и малым шагом используется величина улучшения $D(\boldsymbol{\lambda}^{n+1})$ и $\hat{D}(\boldsymbol{\lambda}^{n+1})$ относительно $D(\bar{\boldsymbol{\lambda}})$ с отсечением по порогу m_L. Формально, критерием выполнения большого шага является выполнение неравенства $D(\boldsymbol{\lambda}^{n+1}) - D(\bar{\boldsymbol{\lambda}}) \geqslant m_L(\hat{D}(\boldsymbol{\lambda}^{n+1}) - D(\bar{\boldsymbol{\lambda}}))$.

Параметры схемы (4.6) w^n отвечают за удержание текущей точки $\boldsymbol{\lambda}^{n+1}$ недалеко от текущей оценки решения $\bar{\boldsymbol{\lambda}}$ и фактически соответствуют обратной величине к длине шага в субградиентном методе. В работе [52] лучше всего себя проявила следующая схема выбора значений w^n:

$$w^n = P_{[w_{\min}, w_{\max}]} \left(\left(\gamma \frac{A^n - \max_{k=1,\ldots,n} D(\boldsymbol{\lambda}^k)}{\|\partial D(\boldsymbol{\lambda}^n)\|_2} \right)^{-1} \right), \tag{4.7}$$

где $P_{[w_{\min}, w_{\max}]}$ означает операцию проекции на отрезок $[w_{\min}, w_{\max}]$. В случае малого шага значение w^n не изменяется.

Важным практически аспектом, необходимым для практического использования метода пучков, является поддержание малого размера пучка. В работе [52] для этой цели используется две схемы:

1. ограничение размера пучка константой с удалением элементов, которые находятся наиболее далеко от оптимума в текущей точке;
2. поддержание усреднённого пучка (aggregated bundle), согласно методу, предложенному в работе [56].

Итого, метод пучков содержит параметры γ, $w_{\min}$, $w_{\max}$, m_L (в работе [56] обозначено m_r), а также максимальный размер пучка b_s при выборе схемы ограничения размера пучка 1.

В дополнение к этим двум схемам, в экспериментах также использовался комбинированный алгоритм LMBM [45] в реализации от авторов метода[1]. В данной схеме был использован только один параметр, отличный от значения по умолчанию: размер пучка b_s.

Алгоритм BFGS. Алгоритм Бройдена-Флетчера-Гольдфарба-Шанно (BFGS) принадлежит классу квазиньютоновских методов. Метод BFGS является методом первого порядка (требует вычисления только значения функции и её градиента) и хорошо себя зарекомендовал для оптимизации гладких функций как с теоретической, так и с практической точек зрения. Льюис и Овертон [90] исследовали вопрос применимости алгоритма BFGS для негладких задач, делая попытку обосновать хорошие результаты метода на практике, замеченные ранее.

Как и другие квазиньютоновские методы, алгоритм BFGS на каждой итерации выбирает направление шага согласно следующему правилу:

$$\boldsymbol{\lambda}^{n+1} = \boldsymbol{\lambda}^n + \alpha^n S^n \partial D(\boldsymbol{\lambda}^n),$$

где S^n некоторая положительно-определённая матрица, обновляемая на каждой итерации, а α^n – длина шага, которая находится при помощи одномерной оптимизации. В работе [90] предлагается версия алгоритма BFGS, специализированная для задач негладкой оптимизации. Данный алгоритм реализован в библиотеке HANSO[2]. В экспериментах только один параметр данной библиотеки использовался со значением не по умолчанию: максимальный ранг матрицы S^n на каждой шаге – h_r.

Неприменимые методы. Проксимальные методы (п. 5), и методы, основанные на сглаживании функционала (п. 6), не удаётся напрямую применить для максимизации нижней оценки SMR. Причина заключается в том, что для работы этих методов необходимо не только уметь вычислять значение и субградиент оптимизируемой функции, но и некоторые дополнительные выражения, что не удаётся сделать для случая SMR.

[1] http://napsu.karmitsa.fi/lmbm/lmbmu/lmbm-mex.tar.gz
[2] http://www.cs.nyu.edu/overton/software/hanso/

Проксимальные методы требуют возможности вычисления проксимального оператора: $\text{prox}(\boldsymbol{\mu}, \alpha) = \arg\max_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda}) - \alpha\|\boldsymbol{\lambda} - \boldsymbol{\mu}\|^2$, что эквивалентно максимизации функции $D(\boldsymbol{\lambda})$ в некоторой окрестности. В работе [139] это удалось сделать за счёт использования специального вида оптимизируемого функционала.

Савчинский и др. [106] приближают негладкую функцию $D(\boldsymbol{\lambda})$ при помощи гладкой функции $\hat{D}(\boldsymbol{\lambda})$, построенной на основе экспоненциального сглаживания операции минимума лагранжиана по прямым переменным: $\hat{D}(\boldsymbol{\lambda}) = -\rho \log \sum_{\boldsymbol{y}} \exp\left(-L(\boldsymbol{y}, \boldsymbol{\lambda})/\rho\right)$. Задача вычисления данного выражения эквивалентна задаче поиска нормировочной константы распределения Гиббса. В общем случае задача является NP-трудной. В работе [106] рассматривается специальный случай (DD TRW с разбиением на деревья), где нормировочная константа может быть вычислена при помощи алгоритмов передачи сообщений. В случае же SMR решить данную задачу не представляется возможным.

Зак и др. [139] применяют при помощи двойственности по Фенхелю формулируют оптимизируемую функцию используя негладкую функцию $\max(\lambda, 0)$ после чего применяют к ней квадратичное сглаживание $(\lambda - \varepsilon/2)[\lambda \geqslant \varepsilon] + (0.5\lambda^2/\varepsilon)[0 \leqslant \lambda \leqslant \varepsilon]$. Применить данный подход к двойственной функции SMR также не удаётся.

Стохастические субградиентные методы (п. 4) хорошо себя зарекомендовали для многих важных задач (например, алгоритм Pegasos для решения оптимизационной задачи, возникающей в методе опорных векторов [110]). Тем не менее для случая SMR методы данной группы не подходят, поскольку предполагают наличие большого количества слагаемых в оптимизируемой функции (для того чтобы на каждой итерации вычислять оценку субградиента лишь по подмножеству слагаемых). В случае же SMR слагаемой всего одно, в случае SMD число слагаемых равно числу меток.

4.3. Максимизация двойственной функции на основе мин-маргиналов

Одним из способов оптимизации двойственной функции $D(\boldsymbol{\lambda})$ является так называемое усреднение мин-маргиналов. Этот подход был применен Уэйнрайтом и др. [130], Колмогоровым [65] для алгоритмов, основанных на разбиении на деревья (TRW), а позднее использован и другими авторами для других разбиений [34, 16]. В данном разделе описано применение подхода, основанного на мин-маргиналах, для алгоритма SMD [97], а также рассмотрена возможность его обобщения на случай SMR.

Определение 9. Мин-маргиналом *функции дискретных переменных* $F(\boldsymbol{z})$, $\boldsymbol{z} = \{z_i\}_{i=1}^N$, $z_i = 1, \ldots, K$, *соответствующим переменной* z_i *и значению* k, *назовём минимальное значение функции* $F(\boldsymbol{z})$ *достигаемое на конфигурациях, в которых переменная* z_i *принимает значение* k:

$$MM_{i,k}F(\boldsymbol{z}) = \min_{\boldsymbol{z}:\, z_i=k} F(\boldsymbol{z}).$$

Заметим, что минимум всех мин-маргиналов, отвечающих одной переменной, равен глобальному минимуму энергии:

$$\min_k MM_{i,k}F(\boldsymbol{z}) = \min_{\boldsymbol{z}} F(\boldsymbol{z}).$$

Одним из частных случаев, допускающих эффективное вычисление мин-маргиналов, является ситуация, когда фактор-граф, задающий зависимости между переменными, не содержит циклов (мин-маргиналы можно легко вычислить по сообщениям, см. раздел 1.3.1.1). В работах [130, 65] используется такой подход.

В отношении алгоритма SMD больший интерес представляет другой частный случай. А именно, если переменные $\boldsymbol{z}$ бинарны и функция $F(\boldsymbol{z})$ парно-сепарабельна и субмодулярна, то все мин-маргиналы можно эффективно вычислить при помощи алгоритма [62], основанного на динамическом построении разрезов похожих графов [59].

Изначально метод усреднения мин-маргиналов создавался для ситуации, когда энергия $E(\boldsymbol{x})$ разбивается на слагаемые (аналогично разделу 1.3.2.2), а потом осуществляется согласование решений подзадач. В этом случае согласование подзадач заключается в изменении потенциалов таким образом, чтобы в разных подзадачах мин-маргиналы были одинаковы. В случае SMD применить такой метод напрямую не удаётся, поскольку релаксируется не условие равенства нескольких переменных, а условие равенства суммы переменных значению 1. В этом случае модифицировать потенциалы при помощи двойственной переменной λ_i надо так, чтобы ровно для одной метки мин-маргинал за 1 был меньше, чем мин-маргинал за 0, а для всех остальных меток мин-маргинал за 0 был меньше, чем мин-маргинал за 1.

Приведём формальное описание метода и результаты, связанные с ним.

Рассмотрим текущее значение двойственных переменных $\boldsymbol{\lambda}^{old}$. Рассмотрим все мин-маргиналы по переменным $\boldsymbol{y}$ лагранжиана (2.7) в точке $\boldsymbol{\lambda}^{old}$: $MM_{ip,k}L\left(\boldsymbol{y}, \boldsymbol{\lambda}^{old}\right) = \min_{\{\boldsymbol{y}\in(1.6),\, |y_{jp}=k\}} L\left(\boldsymbol{y}, \boldsymbol{\lambda}^{old}\right)$, $k \in \{0, 1\}$, $i \in \mathcal{V}$, $p \in \mathcal{P}$.

Выберем некоторую вершину $j \in \mathcal{V}$ и обозначим разность мин-маргиналов за 0 и за 1 через δ_p^j: $\delta_p^j = MM_{jp,0}L\left(\boldsymbol{y}, \boldsymbol{\lambda}^{old}\right) - MM_{jp,1}L(\boldsymbol{y}, \boldsymbol{\lambda}^{old})$, $p \in \mathcal{P}$. Пусть $\delta_{(1)}^j$ – максимальное из чисел δ_p^j, $p \in \mathcal{P}$, $\delta_{(2)}^j$ – следующее наибольшее число (второй максимум); $p_j^{(1)}$ и $p_j^{(2)}$ – соответствующие индексы (аргмаксимумы): $p_j^{(1)} = \arg\max_{p\in\mathcal{P}} \delta_p^j$, $\delta_{(1)}^j = \delta_{p_j^{(1)}}^j$, $p_j^{(2)} = \arg\max_{p\in\mathcal{P}\setminus\{p_j^{(1)}\}} \delta_p^j$, $\delta_{(2)}^j = \delta_{p_j^{(2)}}^j$.

Построим новое значение $\boldsymbol{\lambda}^{new}$ двойственных переменных $\boldsymbol{\lambda}$ по следующему правилу:

$$\lambda_i^{new} = \begin{cases} \lambda_i^{old} + \Delta_j, & i = j, \\ \lambda_i^{old}, & \text{иначе.} \end{cases} \tag{4.8}$$

Здесь i пробегает множество $\mathcal{V}$, а Δ_j – любое число из отрезка (возможно состоящего из одной точки) $\left[\delta^j_{(2)}, \delta^j_{(1)}\right]$.

Далее в данном разделе будет показано (теорема 9), что правило пересчёта (4.8) приводит к неубыванию двойственной функции $D(\boldsymbol{\lambda})$ (2.9), а также, что неподвижная точка такого процесса[3] удовлетворяет условию слабого согласования.

Теорема 9. *Итерационный процесс* (4.8)*, пересчитывающий нижнюю оценку алгоритма SMD* (2.9)*, обладает следующими свойствами:*

1. *На каждой итерации процесса* (4.8) *нижняя оценка* $D(\boldsymbol{\lambda})$ *не убывает.*
2. *Если для вершины* $j \in \mathcal{V}$ *либо* $\delta^j_{(1)} < 0$ *или* $\delta^j_{(2)} > 0$*, то применение пересчёта* (4.8) *приведет к строгому возрастанию нижней оценки.*
3. *Неподвижная точка процесса* (4.8) *удовлетворяет условию слабого согласования.*
4. *Если для вершины* $j \in \mathcal{V}$ *значение* 0 *принадлежит отрезку* $\left[\delta^j_{(2)}, \delta^j_{(1)}\right]$*, то данная точка является максимумом нижней оценки по переменной* λ_j*.*

Доказательство. Перед доказательством основного утверждения теоремы докажем одну вспомогательную лемму.

Лемма 6. *Рассмотрим псевдо-булеву функцию* $F : \mathbb{B}^n \to \mathbb{R}$*. Обозначим за* δ_i *разность между мин-маргиналами* i*-й переменной за* 0 *и за* 1*:* $\delta_i = MM_{i,0}F - MM_{i,1}F$*. Тогда минимум энергии после добавления унарного потенциала* i*-й переменной можно вычислить в явном виде:*

$$\min_{\boldsymbol{z}\in\mathbb{B}^n} \left(F(\boldsymbol{z}) + \delta z_i\right) = \begin{cases} MM_{i,0}F, & \delta \geqslant \delta_i, \\ MM_{i,1}F + \delta, & \delta \leqslant \delta_i. \end{cases}$$

Доказательство. По определению мин-маргинала функции F выполнено $MM_{i,0}F = \min_{\boldsymbol{z}:\, z_i=0} F(\boldsymbol{z})$ и $MM_{i,1}F = \min_{\boldsymbol{z}:\, z_i=1} F(\boldsymbol{z})$. Исходя из этих равенств, можно в явном виде вычислить мин-маргиналы функции $F\boldsymbol{z} + \delta z_i$ для переменной z_i: $MM_{i,0}(F + \delta z_i) = \min_{\boldsymbol{z}:\, z_i=0} F(\boldsymbol{z})$ и $MM_{i,1}(F + \delta z_i) = \min_{\boldsymbol{z}:\, z_i=1} F(\boldsymbol{z}) + \delta$. Доказательство леммы завершает равенство минимума

[3] Неподвижной точкой процесса (4.8) будет называть точку $\boldsymbol{\lambda}$, в которой для всех вершин $j \in \mathcal{V}$ значение 0 принадлежит отрезку $\left[\delta^j_{(2)}, \delta^j_{(1)}\right]$.

энергии минимуму всех мин-маргиналов по одной переменной. Заметим, что в случае равенства $\delta = \delta_i$ утверждение леммы корректно, поскольку выполнено $MM_{i,0} = MM_{i,1}F + \delta$. □

Проведём доказательство теоремы 9. Будем использовать факт того, что в алгоритме SMD лагранжиан распадается на сумму слагаемых:

$$L(\boldsymbol{y}, \boldsymbol{\lambda}) = \sum_{p \in \mathcal{P}} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}) - \sum_{i \in \mathcal{V}} \lambda_i,$$

где слагаемые $\Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda})$ являются субмодулярными функциями переменных $\boldsymbol{y}_p = \{y_{ip}\}_{i \in \mathcal{V}}$ (2.10).

Сначала докажем пункты 1 и 2. Рассмотрим шаг пересчёта при выделенной вершине $j \in \mathcal{V}$. В этом случае, согласно (4.8), значение λ_j изменяется, а все λ_i, $i \neq j$ не изменяются. Рассмотрим отдельно случаи положительного и отрицательного значения Δ_j (если $\Delta_j = 0$, то оба пункта выполнены автоматически).

Пусть $\Delta_j > 0$. Тогда, поскольку $\delta^j_{(1)} = MM_{j,0}\ \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{old}) - MM_{j,1}\ \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{old}) \geqslant \Delta_j > 0$, по лемме 6 выполнено равенство

$$\min_{\boldsymbol{y}_{p_j^{(1)}}} \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{new}) = MM_{j,1}\ \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{old}) + \Delta_j = \min_{\boldsymbol{y}_{p_j^{(1)}}} \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{old}) + \Delta_j. \tag{4.9}$$

Для других слагаемых лагранжиана (2.7) выполнены неравенства (т.к. $\Delta_j > 0$)

$$\min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{new}) \geq \min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}), \quad \forall p \neq p_j^{(1)}. \tag{4.10}$$

При этом, если $\delta^j_{(2)} > 0$, то

$$\min_{\boldsymbol{y}_{p_j^{(2)}}} \Phi_{p_j^{(2)}}(\boldsymbol{y}_{p_j^{(2)}}, \boldsymbol{\lambda}^{new}) = MM_{j,0}\ \Phi_{p_j^{(2)}}(\boldsymbol{y}_{p_j^{(2)}}, \boldsymbol{\lambda}^{old}) > MM_{j,1}\ \Phi_{p_j^{(2)}}(\boldsymbol{y}_{p_j^{(2)}}, \boldsymbol{\lambda}^{old}) = \min_{\boldsymbol{y}_{p_j^{(2)}}} \Phi_{p_j^{(2)}}(\boldsymbol{y}_{p_j^{(2)}}, \boldsymbol{\lambda}^{old}). \tag{4.11}$$

Используя равенство (4.9) и неравенства (4.10), получим

$$D(\boldsymbol{\lambda}^{new}) = \sum_{p \in \mathcal{P}} \min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{new}) - \sum_{j \in \mathcal{V}} \lambda_j^{new} \geq D(\boldsymbol{\lambda}^{old}) + \Delta_j - \Delta_j = D(\boldsymbol{\lambda}^{old}). \tag{4.12}$$

Если ли же $\delta^j_{(2)} > 0$, то, согласно (4.11), неравенство (4.12) выполнено строго.

Рассмотрим случай $\Delta_j < 0$. Тогда по лемме 6 (т.к. $\delta^j_p = MM_{j,0}\ \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) - MM_{j,1}\ \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) \leqslant \Delta_j < 0$, $p \neq p_j^{(1)}$) выполнены равенства

$$\min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{new}) = MM_{j,0}\ \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) = \min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}), \quad \forall p \neq p_j^{(1)} \tag{4.13}$$

и равенство

$$\min_{\boldsymbol{y}_{p_j^{(1)}}} \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{new}) = MM_{j,1}\ \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{old}) + \Delta_j = MM_{j,0}\ \Phi_{p_j^{(1)}}(\boldsymbol{y}_{p_j^{(1)}}, \boldsymbol{\lambda}^{old}) - \delta^j_{p_j^{(1)}} + \Delta_j. \tag{4.14}$$

Если $\delta^j_{(1)} < 0$, то

$$\min_{\boldsymbol{y}_{p^{(1)}_j}} \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{new}) > MM_{j,0}\, \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{old}) + \Delta_j = \min_{\boldsymbol{y}_{p^{(1)}_j}} \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{old}) + \Delta_j. \tag{4.15}$$

Если $\delta^j_{(1)} \geqslant 0$, то

$$\min_{\boldsymbol{y}_{p^{(1)}_j}} \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{new}) = MM_{j,1}\, \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{old}) + \Delta_j = \min_{\boldsymbol{y}_{p^{(1)}_j}} \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{old}) + \Delta_j. \tag{4.16}$$

Из равенств (4.14), (4.16) и неравенства (4.15) следует неравенство (4.12), причём в случае $\delta^j_{(1)} < 0$ неравенство (4.12) выполнено строго. Таким образом, пункты 1 и 2 теоремы 9 доказаны.

Докажем пункт 3. Если $\boldsymbol{\lambda}$ – неподвижная точка, то для всех вершин $j \in \mathcal{V}$, выполнены неравенства $\delta^j_p \leqslant 0 \leqslant \delta^j_{(1)}$, $p \in \mathcal{P} \setminus \{p^{(1)}_j\}$. Отсюда следует, что $1 \in Z_{jp^{(1)}_j}(\boldsymbol{\lambda})$ и $0 \in Z_{jp}(\boldsymbol{\lambda})$, $p \in \mathcal{P} \setminus \{p^{(1)}_j\}$, а значит условие слабого согласования выполнено по определению.

Докажем пункт 4. Пусть для точки $\boldsymbol{\lambda}^{old}$ и вершины $j \in \mathcal{V}$ выполнено условие $0 \in \left[\delta^j_{(2)}, \delta^j_{(1)}\right]$. Рассмотрим разные значения возмущения переменной $\lambda_j - \Delta_j$, и точку $\boldsymbol{\lambda}^{new}$, полученную при помощи возмущения j-й компоненты $\boldsymbol{\lambda}^{old}$ (4.8).

Если $\Delta_j \in \left[\delta^j_{(2)}, \delta^j_{(1)}\right]$, то по лемме 6 выполнены равенства

$$D(\boldsymbol{\lambda}^{new}) = \sum_{p \in \mathcal{P} \setminus \{\delta^j_{(1)}\}} \min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{new}) + \min_{\boldsymbol{y}_{p^{(1)}_j}} \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{new}) - \sum_{j \in \mathcal{V}} \lambda^{old}_j - \Delta_j =$$

$$\sum_{p \in \mathcal{P} \setminus \{\delta^j_{(1)}\}} MM_{j,0}\, \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) + MM_{j,1}\, \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) - \sum_{j \in \mathcal{V}} \lambda^{old}_j = D(\boldsymbol{\lambda}^{old}). \tag{4.17}$$

Если $\Delta_j > \delta^j_{(1)} > 0$, то выполнено неравенство

$$D(\boldsymbol{\lambda}^{new}) = \sum_{p \in \mathcal{P}} MM_{j,0}\, \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) - \sum_{j \in \mathcal{V}} \lambda^{old}_j - \Delta_j =$$

$$\sum_{p \in \mathcal{P} \setminus \{\delta^j_{(1)}\}} \min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) + \min_{\boldsymbol{y}_{p^{(1)}_j}} \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{old}) + \delta^j_{(1)} - \sum_{j \in \mathcal{V}} \lambda^{old}_j - \Delta_j < D(\boldsymbol{\lambda}^{old}). \tag{4.18}$$

В случае $\Delta_j < \delta^j_{(2)} < 0$ получаем аналогичное неравенство:

$$D(\boldsymbol{\lambda}^{new}) = \sum_{p \in \mathcal{P}:\, \delta^j_p < \Delta_j} MM_{j,0}\, \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) + \sum_{p \in \mathcal{P}:\, \delta^j_p \geqslant \Delta_j} \left(MM_{j,1}\, \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) + \Delta_j\right) - \sum_{j \in \mathcal{V}} \lambda^{old}_j - \Delta_j =$$

$$\sum_{p \in \mathcal{P} \setminus \{p^{(1)}_j\}:\, \delta^j_p < \Delta_j} \min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) + \sum_{p \in \mathcal{P} \setminus \{p^{(1)}_j\}:\, \delta^j_p \geqslant \Delta_j} \left(\min_{\boldsymbol{y}_p} \Phi_p(\boldsymbol{y}_p, \boldsymbol{\lambda}^{old}) - \delta^j_p + \Delta_j\right) +$$

$$\min_{\boldsymbol{y}_{p^{(1)}_j}} \Phi_{p^{(1)}_j}(\boldsymbol{y}_{p^{(1)}_j}, \boldsymbol{\lambda}^{old}) + \Delta_j - \sum_{j \in \mathcal{V}} \lambda^{old}_j - \Delta_j < D(\boldsymbol{\lambda}^{old}). \tag{4.19}$$

Последнее неравенство выполнено строго, поскольку множество слагаемых второй группы содержит $p^{(2)}_j$ и $\Delta_j < \delta^j_{(2)}$. Равенство (4.17) и неравенства (4.18) и (4.19) доказывают пункт 4 теоремы 9. □

Вход: лагранжиан $L(\boldsymbol{y}, \boldsymbol{\lambda})$ (2.7), начальное приближение $\boldsymbol{\lambda}_0$;

Выход: покоординатный максимум $\boldsymbol{\lambda}$;

$\boldsymbol{\lambda} := \boldsymbol{\lambda}_0$;
повторять
  converged := true;
  для всех $i \in \mathcal{V}$
    для всех $p \in \mathcal{P}$
      $\delta_p^i := MM_{ip,0}\, L(\boldsymbol{y}, \boldsymbol{\lambda}) - MM_{ip,1}\, L(\boldsymbol{y}, \boldsymbol{\lambda})$;
    если $\delta_{(1)}^i < 0$ **или** $\delta_{(2)}^i > 0$ **то**
      $\lambda_i := \lambda_i + 0.5(\delta_{(1)}^i + \delta_{(2)}^i)$;
      converged := false;
пока converged $\neq$ true;

Алгоритм 1: Алгоритм покоординатного подъёма максимизации нижней оценки $D(\boldsymbol{\lambda})$ SMD (2.9).

У теоремы 9 есть очевидное следствие, позволяющее провести параллели между методом маргиналов для SMD и другими методами максимизации нижних оценок.

Следствие 1. *Неподвижная точка процесса* (4.8) *является покоординатным максимумом нижней оценки* $D(\boldsymbol{\lambda})$.

Это утверждение означает, что процесс (4.8) является алгоритмом покоординатного подъёма, аналогично MSD [8, 5, 132] и простейшему варианту MPLP [117] для стандартной линейной релаксации.

Теорема 9 позволяет построить алгоритм покоординатного подъёма для максимизации нижней оценки SMD (2.9): алгоритм 1. На практике, обычно данный алгоритм позволяет найти покоординатный максимум за 1-2 итерации. Использование динамических разрезов графов [59] при вычислении мин-маргиналов позволяет реализовать алгоритм 1 существенно более эффективно. Серьезным недостатком алгоритма 1 является существенная зависимость точки сходимости от начального приближения, вызванная наличием большого числа покоординатных максимумов нижней оценки (2.9). Достоинством алгоритма 1 является гарантия получения точки, удовлетворяющей условию слабого согласования, за конечное число шагов. Это свойство не выполнено, например, для субградиентного алгоритма [73], который может начать осциллировать в окрестности оптимума.

Обобщение на случай SMR. Формулу пересчёта двойственных переменных $\boldsymbol{\lambda}$ (4.8) можно в явном виде применить и для обобщения SMD – алгоритма SMR (2.16). Тем не менее теорема 9, вообще говоря, не выполнена – данная процедура может приводить к уменьшению нижней оценки. В доказательстве теоремы 9, приведённом выше, существенно используется тот факт, что лагранжиан SMD можно разбить на подзадачи, в каждой из которых пересчёт (4.8) изменяет только один унарный потенциал (не верно для SMR).

Приведём пример, когда формула пересчёта уменьшает значение двойственной функции. Пусть множество вершин $\mathcal{V}$ состоит из одной переменной, а множество меток $\mathcal{P}$ – из трёх: $\mathcal{P} = \{0, 1, 2\}$. Рассмотрим задачу минимизации функции

$$f(\boldsymbol{y}) = 1.5y_0 + 1.5y_1 + 1.5y_2 - y_0y_1 - y_1y_2 - y_0y_2$$

бинарных переменных $\boldsymbol{y}$ при ограничении $y_0 + y_1 + y_2 = 1$. Выпишем двойственную функцию аналитически: $D(\lambda) = \min(-1.5 - \lambda, 2\lambda)$. Пусть текущее значение переменной λ равно 0: $\lambda^{old} = 0$. Тогда все мин-маргиналы за 0 равны -1.5, за 1 – 0. Разности мин-маргиналов δ_0, δ_1, δ_2 все равны -1.5, а значит $\Delta_1 = -1.5$. При этом $D(0) = -1.5$, а $D(-1.5) = -3$. Таким образом, шаг, описываемый уравнением (4.8), приводит к уменьшению нижней оценки, что невозможно в случае SMD.

Несмотря на то, что аналитические формулы пересчёта для SMR выписать не удаётся, эффективно вычислять покоординатный максимум можно при помощи сведения к задаче параметрического максимального потока (parametric max-flow) [69], что, в принципе, позволяет использовать метод оптимизации, основанный на мин-маргиналах и для случая SMR.

4.4. Построение решения прямой задачи

В предыдущих разделах данной главы рассматривалась задача нахождения максимума нижней оценки $D(\boldsymbol{\lambda})$ глобального минимума энергии $E(\boldsymbol{x})$. Обычно, в дополнение к нижней оценке, требуется найти ещё и оценку на прямое решение задачи минимизации энергии.

В литературе встречается две разные постановки задачи построения прямого решения или хотя бы допустимой точки прямой задачи:

1. построение решения релаксации прямой задачи: найти значения переменных являющиеся решением релаксации исходной целочисленной задачи (например, стандартной линейной релаксации при использовании методов DD TRW и SMD);
2. построение решения исходной целочисленной задачи, а именно поиск разметки $\hat{\boldsymbol{x}} \in \mathcal{X}$.

Решение задачи 1 полезно для построения теоретически обоснованного критерия останова для методов оптимизации двойственных функций. Решение задачи 2 необходимо как решение именно исходной дискретной задачи минимизации энергии (в некоторых приложениях никакие релаксации неприемлемы в качестве решения).

В разделе 4.4.1 рассматривается подход к решению задачи 1, применимый для субградиентных методов оптимизации двойственной функции. В последующих разделах рассматриваются вопросы, связанные с решением задачи 2: в разделе 4.4.2 – частичная оптимальность, в разделе 4.4.3 – построение решения при отсутствии зазора, в разделе 4.4.4 – эвристический алгоритм построения решения в общем случае.

4.4.1. Построение целостного дробного решения

Задача построения целостного дробного решения фактически является задачей построения решения оптимизационной задачи с ограничениями по множителям Лагранжа. Результаты, связанные с решением этой задачи, существуют с конца 70-х годов. Например, Н. З. Шор в своей книге [10] описывает метод построения прямого решения для задачи линейного программирования при решении двойственной задачи при помощи метода субградиентного спуска.

В контексте задачи минимизации энергии особняком стоит работа Вернера [134], в которой приводится задача построения прямого решения сводится к задаче выполнимости. Данный метод применим только, если точка максимума двойственной функции известна точно (не применим на ранних итерациях методов оптимизации), но зато не требует знания никакой предыстории оптимизации. Большинство же методов построения прямого дробного решения основано на использовании истории оптимизации и специфики используемого метода (как в [10]). В работе [117, п. 1.7.1] приводится схема, применимая при использовании метода субградиентного спуска для решения двойственной задачи. Работы [106] и [139] анализируют применимость подобных схем для методов, основанных на сглаживании двойственных функций. Савчинский и Шмидт [105] приводят интересные теоретические результаты, позволяющие строить допустимое решение по, вообще говоря, недопустимым.

Далее в данном разделе будет приведен алгоритм построения прямого решения для методов SMD, SMR, и NSMR, применимый при оптимизации их нижних оценок при помощи алгоритмов субградиентного подъёма.

Построение сходящейся последовательности без условия допустимости. Для всех рассматриваемых схем показаны задачи линейного программирования, решения которых по зна-

чению совпадают с максимумами соответствующих нижних оценок (см. главу 3). Рассмотрим результаты из работы [13], применимые в рассматриваемых случаях.

Во время работы алгоритма субградиентного подъёма (4.1) вычисляется последовательность пар точек $\{(\boldsymbol{y}^n, \boldsymbol{\lambda}^n)\}_{n=1}^{\infty}$, где $\boldsymbol{y}^n = \arg\min_{\boldsymbol{y}\in\mathcal{Y}} L(\boldsymbol{y}, \boldsymbol{\lambda}^n)$. При длине шага α^n, удовлетворяющей условиям (4.2), последовательной $\{\boldsymbol{\lambda}^n\}$ сходится к решению двойственной задачи $\boldsymbol{\lambda}^*$. Последовательность $\{\boldsymbol{y}^n\}$, вообще говоря, не сойдется к точке из множества оптимальных решений $\mathcal{Y}^*$ и, более того, может не содержать предельных точек из него. Существует целый ряд схем построения «агрегированной» последовательности $\hat{\boldsymbol{y}}^n$, предельные точки которой принадлежат оптимальному множеству $\mathcal{Y}^*$. Рассмотрим три схемы:

$$\hat{\boldsymbol{y}}^n = \sum_{i=1}^{n} \hat{\alpha}_i^n \boldsymbol{y}^i, \quad \text{где} \quad \hat{\alpha}^i = \frac{\alpha^i}{\sum_{j=1}^{n} \alpha^j}, \ , i = 1, \ldots, n; \tag{4.20}$$

$$\hat{\boldsymbol{y}}^n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}^i; \tag{4.21}$$

$$\hat{\boldsymbol{y}}^n = \rho \boldsymbol{y}^n + (1-\rho)\hat{\boldsymbol{y}}^{n-1}, \ n > 1, 0 < \rho < 1, \quad \hat{\boldsymbol{y}}^1 = \boldsymbol{y}^1. \tag{4.22}$$

Все три схемы представляют собой различные способы усреднить элементы последовательности $\{\boldsymbol{y}^n\}$. При этом схемы отличаются способом распределения весов: в схеме (4.20) наибольшим весом обладают самые старые элементы, в схема (4.21) все веса одинаковы, в схеме (4.22) наибольшими весами обладают самые новые элементы. Интуитивно кажется, что схемы, где новые элементы ценятся больше, должны сходиться быстрее. В работе [13] приводятся следующие факты о работе этих схем:

Утверждение 10. *Пусть в субградиентном методе* (4.1) *с последовательностью длин шагов удовлетворяющей ограничениям* (4.2) *усреднённая последовательность* $\hat{\boldsymbol{y}}^n$ *вычисляется по правилу* (4.20)*. Тогда все предельные точки последовательности* $\hat{\boldsymbol{y}}^n$ *принадлежат оптимальному множеству* $\mathcal{Y}^*$.

Утверждение 11. *Пусть в субградиентном методе* (4.1) *с последовательностью длин шагов* (4.3) *усреднённая последовательность* $\hat{\boldsymbol{y}}^n$ *вычисляется по правилу* (4.21)*. Тогда все предельные точки последовательности* $\hat{\boldsymbol{y}}^n$ *принадлежат оптимальному множеству* $\mathcal{Y}^*$.

Алгоритм, использующий правило (4.22), основан на длине шагов (4.5) и требует алгоритмических изменений относительно (4.1) для обеспечения сходимости. Обзор метода приведен в работе [13], а доказательство в работе [15].

Утв. 10 было описано Шором [10, стр. 151–152]. Утв. 11 получено в работе [112] в немного более общем виде.

Построение допустимой сходящейся последовательности. В работе [105] предложен способ построения последовательности допустимых точек $\tilde{\boldsymbol{y}}^n$, сходящейся к точке, входящей в оптимальное множество $\mathcal{Y}^*$, по сходящейся к некоторой, вообще говоря другой, точке множества $\mathcal{Y}^*$ последовательности точек $\hat{\boldsymbol{y}}^n$, которые, вообще говоря, не являются допустимыми.

Основным понятием в дальнейшем анализе является понятие *оптимизирующей проекции* (optimizing projection) [105]. Пусть $f : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ – непрерывная выпуклая функция двух переменных, $\mathcal{Y} \subseteq \mathbb{R}^n$, $\mathcal{Z} \subseteq \mathbb{R}^m$, $\mathcal{D}$ - замкнутое выпуклое подмножество $\mathcal{Y} \times \mathcal{Z}$, $\mathcal{D}_{\mathcal{Y}} = \{\boldsymbol{y} \in \mathcal{Y} \mid \exists \boldsymbol{z} \in \mathcal{Z} : (\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{D}\}$, а $\Pi_{\mathcal{D}_{\mathcal{Y}}} : \mathbb{R}^n \to \mathcal{D}_{\mathcal{Y}}$ – оператор евклидовой проекции на множество $\mathcal{D}_{\mathcal{Y}}$.

Определение 10. Оптимизирующей проекцией *назовём отображение* $P_{f,D} : \mathcal{Y} \times \mathcal{Z} \to \mathcal{D}$, *такое что* $P_{f,D}((\boldsymbol{y}, \boldsymbol{z})) = (\boldsymbol{y}', \boldsymbol{z}') \in \mathcal{D}$ *тогда и только тогда, когда*

$$\boldsymbol{y}' = \Pi_{\mathcal{D}_{\mathcal{Y}}}(\boldsymbol{y}), \tag{4.23}$$

$$\boldsymbol{z}' = \arg \min_{\boldsymbol{z}:(\boldsymbol{y}',\boldsymbol{z}) \in \mathcal{D}} f(\boldsymbol{y}', \boldsymbol{z}). \tag{4.24}$$

Основным результатом работы [105] является следующее утверждение:

Утверждение 12. *Пусть выпуклый функционал* $f : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ *удовлетворяет условию Липшица. Тогда сходимость последовательности* $(\boldsymbol{y}^n, \boldsymbol{z}^n) \in \mathcal{Y} \times \mathcal{Z}$, $n = 1, \ldots, \infty$ *к точке минимума функции* f *на множестве* $\mathcal{D}$ *влечёт сходимость последовательности оптимизирующих проекций* $P_{f,D}((\boldsymbol{y}^n, \boldsymbol{z}^n))$, $n = 1, \ldots, \infty$ *к (возможной другой) оптимальной точке.*

Заметим, что линейные релаксации, соответствующие всем трём методам SMD, SMR, и NSMR можно записать в следующем виде:

$$\min_{\boldsymbol{y}_i, \boldsymbol{z}} \quad \sum_{i \in \mathcal{V}} \boldsymbol{\theta}_i^{\mathsf{T}} \boldsymbol{y}_i + \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{z} \tag{4.25}$$

$$\text{s.t.} \quad \boldsymbol{y}_i \in \Delta(|\mathcal{P}|),\ i \in \mathcal{V}, \tag{4.26}$$

$$A\boldsymbol{z} \leqslant B\boldsymbol{y} + \boldsymbol{c}. \tag{4.27}$$

Здесь $\Delta(|\mathcal{P}|)$ – стандартный симплекс в пространстве размерности, равной количеству меток $|\mathcal{P}|$, $\boldsymbol{y}_i = \{y_{ip}\}_{p \in \mathcal{P}}$, $\boldsymbol{z}$ – все остальные переменные, входящие в соответствующую задачу линейного программирования, A, B, $\boldsymbol{c}$ матрицы и вектор, определяемые конкретной формулировкой задачи.

Результаты, описанные в предыдущем параграфе говорят, что при использовании метода субградиентного подъёма для максимизации двойственной функции, выборе правильной схемы выбора длины шага и правильной схемы получения «агрегированных» последовательностей $\{(\hat{\boldsymbol{y}}^n, \hat{\boldsymbol{z}}^n)\}$, последовательность $\{(\hat{\boldsymbol{y}}^n, \hat{\boldsymbol{z}}^n)\}$ будет содержать предельную точку, являющуюся решением задачи линейного программирования, равной исходной задачи. Согласно утв. 12 для

построения сходящейся последовательности допустимых точке нужно для каждой точки $(\hat{\boldsymbol{y}}^n, \hat{\boldsymbol{z}}^n)$ вычислить оптимизирующую проекцию. Для этого нужно

1. каждый вектор $\hat{\boldsymbol{y}}_i^n$, $i \in \mathcal{V}$ ортогонально спроектировать на единичный симплекс (4.26),
2. минимизировать целевую функцию относительно оставшихся переменных $\boldsymbol{z}$ при фиксированных значениях $\hat{\boldsymbol{y}}^n$.

Заметим, что значения $\hat{\boldsymbol{z}}^n$ нигде не используются, а значит их можно не вычислять.

Для проектирования точки на единичный симплекс существует много эффективных алгоритмов, например, алгоритм [28] требует одной сортировки и одного прохода по точкам.

Второй шаг вычислительно более сложен и свой для каждого из методов:

SMD находит нижнюю оценку равную значению стандартной линейной релаксации (опр. 5), также как и DD TRW. Как было замечено в работах [106, 105] при фиксации унарных переменных y_{ip}, $i \in \mathcal{V}$, $p \in \mathcal{P}$ целевая функция (1.23) распадается на отдельные подзадачи (своя для каждого ребра $\{i,j\} \in \mathcal{E}$), каждая из которых является транспортной задачей небольшого размера (можно решить, например, при помощи венгерского алгоритма). Выражение (3.14) для данной нижней оценки, использованной при доказательстве теоремы 1, показывает что переменные $y_{ij,p}$, $\{i,j\} \in \mathcal{E}$, $p \in \mathcal{P}$ можно просто положить равными $\min(y_{ip}, y_{jp})$, дополнив остальные числа, используя конструктивное доказательство леммы 3 [98, лемма 2].

SMR находит нижнюю оценку, равную решению задачи (3.23)-(3.26). Аналогично предыдущему пункту, каждую переменную $y_{C,\boldsymbol{d}}$, $C \in \mathcal{C}$, $\boldsymbol{d} \in \mathcal{D}_C$ можно положить равной $\min_{\ell \in C} y_{\ell d_\ell}$.

NSMR находит нижнюю оценку, равную решению задачи (3.35)-(3.39). При фиксации переменных y_{ip}, $i \in \mathcal{V}$, $p \in \mathcal{P}$ оптимизационная задача распадается на подзадачи (по одной для каждого ребра $\{i,j\} \in \mathcal{E}$) небольшого размера. Каждая из таких подзадач является задачей линейного программирования и может быть решения, например, каким-либо методом внутренних точек.

4.4.2. Частичная оптимальность

Одним из наиболее привлекательных свойств приближённых методов минимизации энергии является, так называемая, частичная оптимальность (partial optimality, persistency). Данное свойство заключается в том, что по итогам работы метода вычисляется «частичная разметка» (для каждой переменной множества $\mathcal{V}$ либо присваивание одной из меток множества $\mathcal{P}$, либо отказ отказ от разметок этой переменной), гарантирующая существование полной оптимальной

разметки всех переменных, совпадающей с частичной разметкой в тех вершинах, где не было отказа.[4]

Наболее известным методом, обладающим свойством частичной оптимальности, является алгоритм QPBO минимизации парно-сепарабельных (вообще говоря, несубмодулярных) энергий бинарных переменных [22, 66]. В последние годы появился целый ряд работ, делающих попытки улучшить качество работы алгоритма QPBO (количество размеченных вершин) [103] и расширить область его применимости [74, 63, 119]. Также было показано, что использование частичной оптимальности (если метод построения частичной разметки работает быстро) может приводить к существенному ускорению процедуры минимизации энергии в целом [11].

Из всей группы декомпозиционных методов что-либо про частичную оптимальность известно лишь про метод TRW, основанный на усреднении мин-маргиналов [130]. Колмогоровым и Уэйнрайтом [67] было показано, что TRW обладает свойством частичной оптимальности, только если в исходной энергии все переменные бинарны. При этом вершина считается размеченной, тогда и только тогда когда в ней все подзадачи согласованы.

В рамках работы [96] был рассмотрен вопрос о выполнении свойства частичной оптимальности для методов группы SMR. На данный вопрос получен отрицательный ответ. Пример энергии, на которой алгоритм SMD (наиболее простой частный случай SMR) не обладает свойством частичной оптимальности, приведен на рис. 4.2. Такой результат аналогичен выводам работы [67] об алгоритмах TRW. Когда исходные переменные бинарны, частичная оптимальность выполнена (т.к. SMD эквивалентно QPBO), иначе не выполнена.

4.4.3. Построение прямого решения при нулевом зазоре

В этом разделе приводится несколько результатов, позволяющих построить оптимальную полную разметку прямой задачи по найденному максимуму нижней оценки $D(\boldsymbol{\lambda})$ (2.9) для случая SMD. Эти результаты (теорема 10 и следствие 2) позволяют построить оптимальное решение исходной задачи в некоторых ситуациях. Условия этих теорем допускают эффективную проверку и фактически расширяют сертификат оптимальности. Если одна из них оказывается применима, то зазор между прямой и двойственной задачей равен 0, но построение решение затруднено тем, что минимум лагранжиана по бинарным переменным $\boldsymbol{y}$ не единственен (в окрестности этой точки выполнена ситуация 2 раздела 4.1.2).

Приведём вспомогательное утверждение, аналогичное теореме 3 из работы [67].

[4] Здесь и далее под частичной оптимальностью понимается слабая частичная оптимальность (weak persistency). В литературе также встречается понятие сильной частичной оптимальности (strong persistency). В этом случае утверждается, что *любая* оптимальная полная разметка совпадает с частичной в вершинах, где не было отказа.

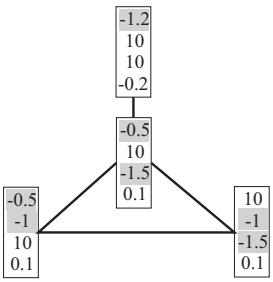


Рисунок 4.2.: Пример энергии, при применении к которой метод SMR не обладает свойством частичной оптимальности. Рассматривается энергия, зависящая от 4-х переменных, каждая из которых может принимать 4 значения. Граф зависимостей и унарные потенциалы изображены на рисунке. Унарные потенциалы выбраны таким образом, чтобы максимум нижней оценки $D(\boldsymbol{\lambda})$ достигался в нулевой точке ($\boldsymbol{\lambda}^* = 0$). Парные потенциалы являются потенциалами Поттса: $\theta_{ij}(x_i, x_j) = 2[x_i \neq x_j]$. Метки k, такие что $\{1\} \subseteq Z_{ik}(\boldsymbol{\lambda}^*)$ выделены серым. В верхней вершины все подзадачи согласованы: ровно одна метка допускает значение 1. Свойство частичной оптимальности не выполнено, поскольку глобальный минимум энергии достигается на разметке, при которой все переменные принимают 4-е (нижнее) значение.

Лемма 7. *Пусть* $\boldsymbol{z}^* = \arg\min_{\boldsymbol{z}} F(\boldsymbol{z})$*, где* $F : \{0,1\}^n \to \mathbb{R}$ *– субмодулярная псевдо-булева функция. Рассмотрим разметку* $\boldsymbol{z}^{(1)}$*, такую что*

$$z_i^{(1)} = \begin{cases} 1, & \text{если } MM_{i,0}F(\boldsymbol{z}) = MM_{i,1}F(\boldsymbol{z}), \\ z_i^*, & \text{иначе}, \end{cases}$$

и разметку $\boldsymbol{z}^{(0)}$*, такую что*

$$z_i^{(0)} = \begin{cases} 0, & \text{если } MM_{i,0}F(\boldsymbol{z}) = MM_{i,1}F(\boldsymbol{z}), \\ z_i^*, & \text{иначе}. \end{cases}$$

Тогда выполнены следующие равенства:

$$F(\boldsymbol{z}^*) = F(\boldsymbol{z}^{(1)}) = F(\boldsymbol{z}^{(0)}).$$

Доказательство. Данное утверждение следует из определения 3 субмодулярности и определения 9 мин-маргиналов. □

Приведём два утверждения, предоставляющие возможность строить разметку глобального оптимума исходной задачи в некоторых частных случаях.

Через $Free(p, \boldsymbol{\lambda})$, $p \in \mathcal{P}$, обозначим подмножество множества вершин $\mathcal{V}$, содержащее все вершины $j \in \mathcal{V}$, такие что $Z_{jp}(\boldsymbol{\lambda}) = \{0,1\}$.

Теорема 10. *Пусть точка* $\boldsymbol{\lambda}^0$ *удовлетворяет условию слабого согласования и* $\boldsymbol{y}^0 = \arg\min_{\boldsymbol{y}\in(1.6)} L(\boldsymbol{y}, \boldsymbol{\lambda}^0)$. *Пусть для некоторой метки* $p \in \mathcal{P}$ *выполнены вложения* $Free(q, \boldsymbol{\lambda}^0) \subseteq Free(p, \boldsymbol{\lambda}^0)$, *где* $q \in \mathcal{P} \setminus \{p\}$. *Тогда разметка*

$$y^*_{jr} = \begin{cases} 1, & j \in Free(r, \boldsymbol{\lambda}^0), \quad r = p; \\ 0, & j \in Free(r, \boldsymbol{\lambda}^0), \quad r \neq p; \\ y^0_{jr}, & \textit{иначе}, \end{cases}$$

является решением задачи (1.5)-(1.7)*:* $\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}\in(1.6),(1.7)} E_I(\boldsymbol{y})$.

Доказательство. Лагранжиан (2.7) является суммой слагаемых (2.10), отвечающих различным меткам $r \in \mathcal{P}$. Все слагаемые являются субмодулярными функциями, а значит к ним применима лемма (7) (условие $j \in Free(r, \boldsymbol{\lambda}^0)$ эквивалентно условию $MM_{j,0}\Phi_r(\boldsymbol{y}_r, \boldsymbol{\lambda}^0) = MM_{j,1}\Phi_r(\boldsymbol{y}_r, \boldsymbol{\lambda}^0)$). Правило, построения разметки $\boldsymbol{y}^*$, описанное в условии данной теоремы, фактически является переходом к разметке из единиц в подзадаче метки p и переходом к разметкам из нулей во всех остальных подзадачах. Поскольку, в дополнение к вышесказанному, точка $\boldsymbol{\lambda}^0$ удовлетворяет условию слабого согласования, то разметка построенная по условию данной теоремы удовлетворяет условиям (1.5)-(1.7) и является агрминимумом лагранжиана (2.7) в точке $\boldsymbol{\lambda}^0$. □

Пусть I – множество всех вершин, которые входят в хотя бы одно множество $Free(p, \boldsymbol{\lambda})$, $p \in \mathcal{P}$. Обозначим через I^k, $k = 1, \ldots, K$ компоненты связности множества I. Для каждой компоненты связности теорему 10 можно применять независимо. Следствие 2 и формализует этот результат.

Следствие 2. *Пусть точка* $\boldsymbol{\lambda}^0$ *удовлетворяет условию слабого согласования и* $\boldsymbol{y}^0 = \arg\min_{\boldsymbol{y}\in(1.6)} L(\boldsymbol{y}, \boldsymbol{\lambda}^0)$. *Если для каждой компоненты связности* I^k *существует метка* p^k, *такая что* $(Free(q, \boldsymbol{\lambda}^0) \cap I^k) \subseteq (Free(p^k, \boldsymbol{\lambda}^0) \cap I^k)$, *где* $q \in \mathcal{P} \setminus \{p^k\}$, *то разметка*

$$y^*_{jr} = \begin{cases} y^0_{jr}, & j \notin I, \\ 1, & r = p^k, \quad j \in I^k, \\ 0, & \textit{иначе}, \end{cases}$$

является решением задачи (1.5)-(1.7).

Обозначим за $Free^k(p, \boldsymbol{\lambda})$ k-ю компоненту связности множества $Free(p, \boldsymbol{\lambda})$.

Следствие 3. *Пусть точка* $\boldsymbol{\lambda}^0$ *удовлетворяет условию слабого согласования и* $\boldsymbol{y}^0 = \arg\min_{\boldsymbol{y}\in(1.6)} L(\boldsymbol{y}, \boldsymbol{\lambda}^0)$. *Если существует разбиение* $Free^{k_1}(p_1, \boldsymbol{\lambda}^0), \ldots, Free^{k_m}(p_m, \boldsymbol{\lambda}^0)$ *множества* I ($Free^{k_i}(p_i, \boldsymbol{\lambda}^0) \cap Free^{k_j}(p_j, \boldsymbol{\lambda}^0) = \varnothing$, $i \neq j$, *и* $\bigcup_i Free^{k_i}(p_i, \boldsymbol{\lambda}^0) = I$), *то разметка*

$$y^*_{jr} = \begin{cases} y^0_{jr}, & j \notin I, \\ 1, & r = p_i,\ i \in Free_{k_i}(p_i, \boldsymbol{\lambda}^0),\ i = 1, \ldots, m, \\ 0, & \textit{иначе}, \end{cases}$$

является решением задачи (1.5)-(1.7).

Следствие 3 доказывается при помощи применения теоремы 10 к каждой компоненте $Free^{k_i}(p_i, \boldsymbol{\lambda}^0)$.

4.4.4. Построение прямого решения в общем случае

При использовании алгоритма SMR на практике часто необходимо строить допустимую точку задачи (1.5)-(1.7), обладающую невысоким (пусть и не оптимальным) значением целевой функции (1.5). Построение таких решений необходимо как после завершения работы SMR (если не удаётся получить сертификат оптимальности), так и во время работы метода для оценки текущего прогресса.

Далее в данном разделе приводится быстрый метод поиска допустимой точки с низким значением энергии. Метод основан на идее блочно-координатной минимизации, используемой как самостоятельный метод минимизации энергии ещё в 80-х гг. – алгоритм ICM [19], и похож на аналогичный метод, используемый для получения прямого решения в алгоритме TRW-S [65]. В рамках данной работы будем называть описанный ниже метод ICM.

Пусть дана некоторая точка $\boldsymbol{\lambda}$ и для неё на основе мин-маргиналов построены множества $Z_{ip}(\boldsymbol{\lambda}) \subseteq \{0, 1\}$, $i \in \mathcal{V}$, $p \in \mathcal{P}$.

Инициализацию текущей целочисленной (1.6) и целостной (1.7) разметки $\boldsymbol{y}$ проведём следующим образом: для каждой вершины $i \in \mathcal{V}$ положим $y_{ip} = 1$ для такой метки $p \in \mathcal{P}$, для которой $\{1\} \subseteq Z_{ip}(\boldsymbol{\lambda})$ (если таковых несколько, то выберем любую из них; если таковых нет, то выберем любую метку) и $y_{iq} = 0$ для всех $q \neq p$. Для каждого потенциала $C \in \mathcal{C}$ определим $\hat{\boldsymbol{y}}_C = \{y_{ip}\}_{i\in C}^{p\in\mathcal{P}}$, как разметку переменных $\boldsymbol{y}_C$, минимизирующую функцию, состоящую из потенциала C и унарных потенциалов, инцидентных ему:

$$\sum_{p\in\mathcal{P}} \sum_{i\in C} (\theta_i(p) + \lambda^*_i) y_{ip} + \sum_{\boldsymbol{d}\in\mathcal{D}_C} \theta_C(\boldsymbol{d}) \prod_{\ell\in C} y_{\ell d_\ell}$$

при ограничениях целочисленности и целостности: $\sum_{p\in\mathcal{P}} y_{ip} = 1$, $y_{ip} \in \{0,1\}$. Попытаемся уменьшить общую целевую функцию (1.5) при помощи присваивания $\boldsymbol{y}_C = \hat{\boldsymbol{y}}_C$. Если значение целевой функции уменьшается, то принимаем данное присваивание, иначе отвергаем его.

Проходов по потенциалам $C \in \mathcal{C}$ можно сделать несколько. На промежуточных итерациях SMR даже 1 проход ICM позволяет существенно улучшить текущее значение энергии. При вычислении окончательного значения энергии можно запустить ICM до сходимости (обычно 3-4 прохода по всем вершинам). Если применение ICM слишком замедляет общий алгоритм, то его можно применять не на каждой итерации, а существенно реже (например, на каждой 10-й итерации).

5. Экспериментальное сравнение

Данная глава посвящена экспериментальному исследованию методов рассматриваемых в данной работе.

Раздел 5.1 содержит эксперименты на реальных данных, позволяющие сравнить работу различных методов оптимизации нижней оценки (см. раздел 4.2) подхода SMR. Проведённый эксперимент предоставляет сравнение подхода SMD с аналогами: DD TRW и TRW-S.

Раздел 5.2 содержит результаты экспериментов о применении SMR к задачам с потенциалами высоких порядков. Раздел 5.3 – о применении NSMR к энергиям с неассоциативными парными потенциалами. В разделе 5.4 приведены эксперименты с глобальными ограничениями на переменные.

5.1. Парно-сепарабельные ассоциативные энергии

В данном разделе описан эксперимент, преследующий две цели:

1. исследование работы различных методов оптимизации двойственной функции (см. раздел 4.2), построенной при помощи варианта SMR для ассоциативных парно-сепарабельных энергий – алгоритма SMD (см. раздел 2.1);
2. сравнить работу алгоритма SMD и известного алгоритма двойственной декомпозиции, основанного на разложении графа энергии на деревья – алгоритма DD TRW [73].

Для достижения цели 1 в рамках данного эксперимента рассматривались следующие методы оптимизации нижней оценки:

1. субградиентный подъём (4.1) с адаптивным размером шага (4.5) [73];
2. метод пучков субградиентов (4.6), (4.7) [52] с ограничение размера пучка (п. 1);
3. метод пучков субградиентов (4.6), (4.7) [52] с усреднением пучка (п. 2) [56];
4. комбинированный метод пучков: алгоритм LMBM [45];
5. негладкий вариант алгоритма BFGS [90];

Субградиентный метод с неадаптивными вариантами выбора длины шага (4.3) и (4.4) сходился очень медленно и был сразу исключен из рассмотрения (не попал в список выше).

Рисунок 5.1.: Изображения, для задач семантической сегментации которых, были построены энергии, использованные в эксперименте, описанном разделе 5.1. Верхняя строка – исходные изображения, средняя срока – правильные ответы (чёрным показаны неразмеченные пиксели), нижняя строка – сегментации, соответствующие глобальным минимумам энергий (везде найдены при помощи метода SMD). Цвета меток классов соответствуют цветам, использованным в базе MSRC [114], не в полной мере.

Сравнение различных схем построения нижних оценок с использованием релаксации Лагранжа (например, SMD и DD TRW) осложнено тем, что работоспособность каждой из схем существенно зависит от метода оптимизации, используемого для максимизации соответствующей нижней оценки. Исходя из этой трудности, для достижения цели 2 для алгоритма DD TRW были применены все методы оптимизации, использованные и для SMD. Сравнение также проводилось относительно алгоритма TRW-S [65], максимизирующего нижнюю оценку при помощи передачи сообщений.

Данные. Эксперименты данного раздела проводились на задачах минимизации энергий, построенных по реальным данным Алахари и др. [11][1]. Были использованы задачи двух видов: семантическая многоклассовая сегментация (5 задач, см. рис. 5.1) и выровненное стереосопоставление (4 задачи, см. рис. 5.2). Графы энергий всех задач представляли собой прямоугольные решётки, где вершины соответствовали пикселям изображений, а рёбра – 4-х связной системе соседства. В задачах семантической сегментации размер решёток составлял примерно 200 на 300 ($|\mathcal{V}| \approx 60000$), количество меток $|\mathcal{P}|$ составляло от 3 до 4. В задачах стерео примерный размер решёток – 400 на 450 ($|\mathcal{V}| \approx 180000$), а количество меток $|\mathcal{P}|$ – от 16 до 60. Для задач стерео в качестве парных потенциалов использовались стандартные потенциалы Поттса с весом

[1] http://www.di.ens.fr/~alahari/data/pami10data.tgz

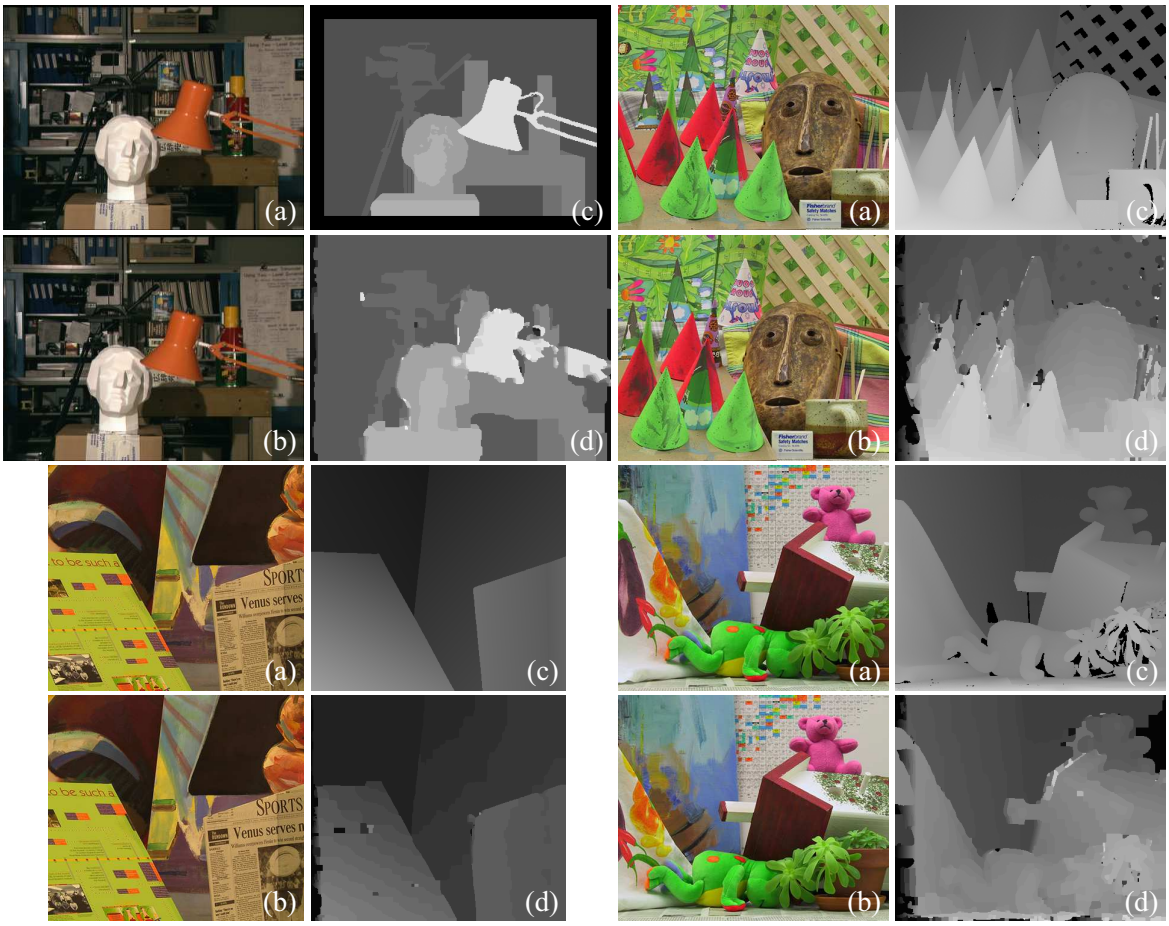


Рисунок 5.2.: Стереопары, для задач стерео-сопоставления которых, были построены энергии, использованные в эксперименте, описанном в разделе 5.1. Изображения (a) – изображения, снятые при помощи левой камеры стереопары, изображения (b) – при помощи правой. (c) – правильные ответы задач, построенные при помощи лазерного сканирования. (d) – результаты, полученные при помощи минимизации энергий [120] алгоритмом SMD.

20, унарные потенциалы были основаны на сопоставлении цветов окрестностей сопоставляемых пикселей. Энергии для задач семантической сегментации были построены при помощи модели TextonBoost [114]: унарные потенциалы совмещали информацию о цвете, положении на изображении и текстуре; парные потенциалы представляли собой взвешенные потенциалы Поттса (вес конкретного потенциала зависел от градиента изображения на соответствующем ребре).

Выбор параметров методов оптимизации. Для каждого исследуемого метода оптимизации, для каждого вида нижней оценки (SMD или DD TRW) и для каждого типа задач (сегментация или стерео) выбор наилучших значений параметров осуществлялся при помощи перебора по сетке значений. Критерием качества набора параметров являлось максимальное значение нижней оценки глобального минимума энергии, найденной за первые 25 секунд работы метода для задач сегментации и за первые 200 секунд для задач стерео. Для метода пучков с ограничением

Таблица 5.1.: Наилучшие значения параметров методов оптимизации нижних оценок, построенных при помощи подходов SMD и DD TRD.

	Метод оптимизации	**Сегментация**	**Стерео**
SMD	субградиентный подъём (4.1), адаптивный размер шага (4.5)	$\gamma = 0.7$	$\gamma = 0.3$
	метод пучков (4.6), (4.7), ограничение размера пучка (п. 1)	γ=0.1, m_L=0.2, $w_{\max}$=100	γ=0.01, m_L=0.1, $w_{\max}$=100
	метод пучков (4.6), (4.7), усреднение пучка (п. 2)	γ=0.02, m_r=0.001, $w_{\max}$=500	γ=0.02, m_r=0.001, $w_{\max}$=1000
DD TRW	субградиентный подъём (4.1), адаптивный размер шага (4.5)	$\gamma = 0.3$	$\gamma = 0.1$
	метод пучков (4.6), (4.7), ограничение размера пучка (п. 1)	γ=0.01, m_L=0.05, $w_{\max}$=1000	γ=0.01, m_L=0.05, $w_{\max}$=2000
	метод пучков (4.6), (4.7), усреднение пучка (п. 2)	γ=0.02, $w_{\max}$=100, m_r=0.0005	γ=0.01, m_r=0.002, $w_{\max}$=500

размера пучка большие размеры пучка приводили к лучшим результатам, но замедляли скорость работы метода. Значение максимального размера пучка $b_s = 100$ было выбрано как компромиссный вариант. Аналогично выбирались значения параметров $b_s = 10$ для метода LMBM и $h_r = 10$ для метода BFGS. Параметр $w_{\min}$ для обеих версий метода пучков не влиял на работу методов и был выставлен 10^{-10} как рекомендовано в работе [52]. Найденные значения всех остальных параметров представлены в таблице 5.1.

Детали реализации. В качестве алгоритмов BFGS и LMBM использованы оригинальные авторские реализации: библиотеки HANSO[2] (реализация в системе MATLAB) и LMBM[3] (реализация на языке Fortran с mex-оболочкой для системы MATLAB). В качестве алгоритмов субградиентного подъёма и двух версий метода пучков использованы собственные реализации в системе MATLAB. Для вычисления значений нижних оценок и их субградиентов (подходы SMD и DD TRW) применялись низкоуровневые реализации на языке C++. Для вычисления нижней оценки SMD использовался алгоритм Бойкова-Колмогорова поиска максимального потока в гра-

[2] http://www.cs.nyu.edu/overton/software/hanso/
[3] http://napsu.karmitsa.fi/lmbm/lmbmu/lmbm-mex.tar.gz

фе, поддерживающий ускоренное построение похожих графов [23, 59] в авторской реализации[4]. Для вычисления нижней оценки DD TRW использовалась собственная реализация алгоритма динамического программирования, специально ускоренная для модели Поттса (работает за линейное, а не квадратичное по числу меток время). В качестве алгоритма TRW-S использовалась оригинальная авторская реализация на языке C++, оптимизированная для работы с моделями Поттса[5].

Все энергии были нормированы таким образом, чтобы энергия результата алгоритма α-расширение равнялась 100, а энергия разметки, полученной как минимум унарных потенциалов, равнялась 0.

Разбиение исходного графа $\mathcal{G}$ на деревья в алгоритме DD TRW строилось таким образом, чтобы каждое ребро графа входило в одно и только одно дерево (для того чтобы сократить количество двойственных переменных). В рамках данного эксперимента использовался наиболее популярный способ построения таких разбиений для графов-решёток с 4-х связностью: разбиение на отдельные столбцы и строки решётки. Такое разбиение использовалось в ряде работ, посвящённых максимизации нижней оценки, построенной DD TRW [106, 52].

Результаты. Результаты проведённого эксперимента приведены на рис. 5.3. Из приведённых графиков можно сделать следующие выводы:

1. все рассмотренные методы оптимизации нижних оценок (как для SMD, так и для DD TRW) при правильном выборе параметров показывают похожие результаты; часть методов менее чувствительна к выбору параметров, часть – более; метод LMBM показал себя наименее чувствительным к выбору параметров и часто достаточно эффективным, а значит может быть рекомендован для использования на практике;
2. алгоритмы, основанные на подходе SMD работают в целом быстрее, чем алгоритмы, основанные на подходе DD TRW; скорее всего, это вызвано тем фактом, что максимизации нижних границ SMR происходят в пространствах намного меньших размерностей;
3. в ситуациях, когда в исходной задаче мало меток, алгоритмы, основанные на SMD, работают быстрее стандартного метода TRW-S; скорее всего, это вызвано тем, что в этом случае ограничения целостности оказывают меньшее влияние на оптимальное решение.

[4] http://pub.ist.ac.at/~vnk/software/maxflow-v3.02.src.tar.gz
[5] http://pub.ist.ac.at/~vnk/papers/TRW-S.html

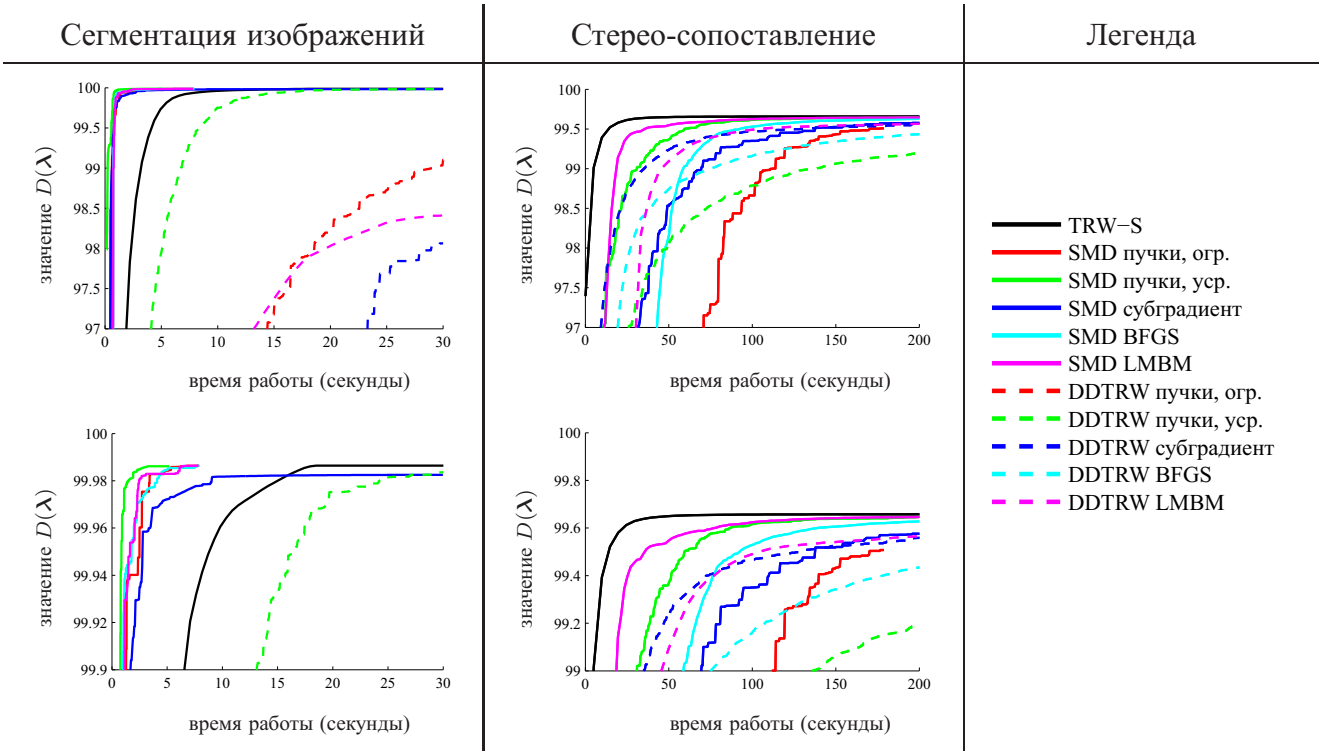


Рисунок 5.3.: Результаты эксперимента, описанного в разделе 5.1. Приведённые графики показывают зависимости нижних оценок глобального оптимума энергии, полученных при помощи различных методов, от времени работы метода. Каждый график является усреднением процесса по нескольким энергиям. Левый столбец соответствует энергиям, построенным для задач сегментации изображений, центральный – энергиям задач стерео-сопоставления. Нижние графики левого и центрального столбцов являются увеличенными версиями соответствующих им верхних. В правом столбце приведена легенда, соотносящаяся названия методов и цвета графиков.

5.2. Энергии с потенциалами высоких порядков

Данный раздел содержит описание экспериментов по оценке применимости подхода SMR для энергий с разреженными потенциалами высоких порядков, а также по сравнению подхода SMR с подходом CWD [71].

Данные. Эксперименты проводятся на двух наборах модельных энергий и одной энергии, построенной для задачи сегментации изображений (данные генерируются аналогично работе [65, раздел 5.1].

Каждый из наборов модельных энергий состоит из 20 энергий. Все энергии зависят от 10-значных переменных, связанных в соответствии с графом вида 4-х связная решётка размера 50 на 50 вершин. Ко всем энергиям добавлены потенциалы высоких порядков, видом которых и отличаются наборы энергий. Все унарные потенциалы за все классы генерируются из нормальных

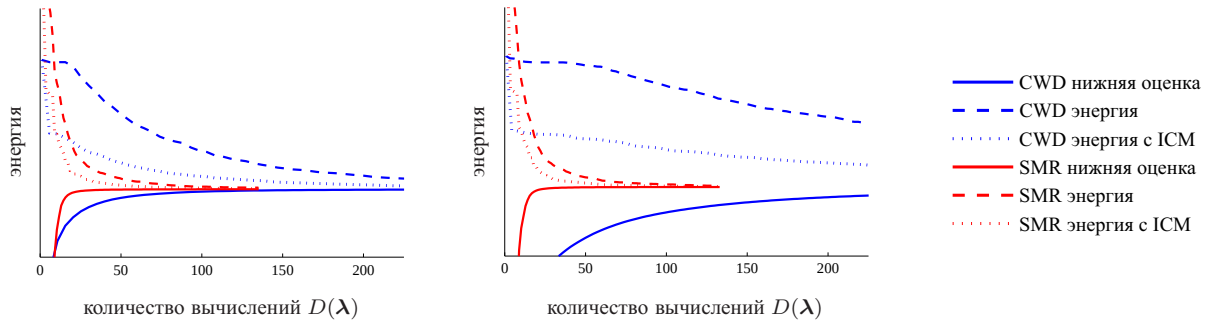
распределений $\mathcal{N}(0, 10)$. Все парные потенциалы являются взвешенными потенциалами Поттса: $0.1|c_{ij}|[x_i \neq x_j]$. Веса c_{ij} генерируются из стандартного нормального распределения $\mathcal{N}(0, 1)$. К каждой из энергий добавляется 50 случайных потенциалов Поттса высокого порядка ($\mathcal{P}^n$-Potts). В первом наборе данных каждый потенциал высокого порядка зависит от 50 вершин, выбранных случайным образом, во втором – от 500. Веса потенциалов высоких порядков генерируются из равномерного распределения на отрезке $[0, 100]$.

Последней рассматриваемой энергией является энергия для задачи сегментации изображения папоротников (рис. 5.5а) на три класса: «папоротник», «земля», «трава». Для каждого класса выбрана небольшая область пикселей этого класса – семена. Пиксели-семена каждого класса рассматриваются как выборка объектов с тремя признаками (цветовое пространство Luv) и по этой выборке восстанавливается плотность распределений, как смесь трёхмерных нормальных распределений из 5 компонент. Унарные потенциалы энергии за каждый из классов строятся как отрицательные логарифмы значений восстановленных плотностей распределений цветов пикселей. Парные потенциалы – потенциалы Поттса с весом 1. Потенциалы высоких порядков – потенциалы Поттса высокого порядка, зависящие от переменных, соответствующих группам пикселей (суперпикселям), построенным при помощи алгоритма EDISON[6] [29] с параметрами по умолчанию. Веса всех потенциалов высоких порядков равны 100.

Детали реализации. Для реализации алгоритма CWD осуществляется следующая декомпозиция энергии: все парные потенциалы распределяются по горизонтальным и вертикальным цепочкам, каждый потенциал высоких порядков относится к отдельной подзадаче, каждый унарный потенциал распределяется равномерно по всем подзадачам, в которых есть зависимость от соответствующей переменной. При такой декомпозиции двойственная функция $D(\boldsymbol{\lambda})$ зависит от $|\mathcal{V}|K(1 + h)$ переменных, где h – среднее количество потенциалов высоких порядков, зависящих от конкретной переменной. Двойственная функция в SMR зависит от $|\mathcal{V}|$ переменных. Для модельных наборов данных двойственные функции CWD зависят от 50000 и 275000 переменных (для первого и второго наборов), двойственная функция SMR зависит от 2500 переменных для обоих наборов. Для энергии сегментации изображения двойственные функции зависят от 675000 (CWD) и 112500 (SMR) переменных.

Для вычисления значения двойственной функции SMR используется алгоритм Бойкова-Колмогорова построения максимального потока в сети [23]; для вычисления значения двойственной функции CWD используется алгоритм динамического программирования для минимизации энергий подзадач, соответствующих столбцам и строкам, и алгоритм, предложенный в рабо-

[6] http://coewww.rutgers.edu/riul/research/code/EDISON/



(a) (b) (c)

Рисунок 5.4.: Энергии и нижние оценки, полученные при помощи максимизации нижних оценок SMR и CWD, усреднённые по наборам модельных энергий. График (a) соответствует энергиям с небольшими потенциалами высоких порядков (порядок 50), график (b) – энергиям с большими потенциалами высоких порядков (порядок 500). Таблица (c) поясняет значение каждой изображенной кривой.

те [71] (см. раздел 1.3.2.2.3), для минимизации энергий подзадач, соответствующих потенциалам высоких порядков. В данном эксперименте вычисления нижних оценок CWD и SMR занимали примерно одинаковое время. Для исключения фактора эффективности реализации время работы измерялось в количестве вычислений функции.

Для максимизации всех нижних оценок всех рассматриваемых энергий использовался негладкий вариант алгоритма BFGS в реализации из библиотеки HANSO. Эксперимент проводился с двумя способами вычисления прямого решения по текущему значению двойственной функции: с использованием алгоритма ICM (см. раздел 4.4.4) и без него.

Результаты. Результаты эксперимента с модельными энергиями приводятся на рис. 5.4. Качественные результаты эксперимента на задаче сегментации энергии приведены на рис. 5.5, количественные – на рис. 5.6. Исходя из этих результатов можно сделать несколько выводов. Во-первых, алгоритм, максимизирующий нижнюю оценку SMR, сходится быстрее, чем алгоритм, максимизирующий нижнюю оценку CWD. Во-вторых, применение алгоритма ICM для получения прямых решений не существенно улучшает текущее значение энергии в случае SMR и существенно – в случае CWD.

Стоит отметить, что улучшенная версия алгоритма CWD – PatB [71], могла бы как-то улучшить результаты CWD только в случае сильно пересекающихся потенциалов высоких порядков (в PatB потенциалы высоких порядков, зависящие от пересекающихся множеств переменных, объединяются в подзадачи специального вида) – график 5.4b. Во всех остальных случаях алгоритм PatB эквивалентен алгоритму CWD.

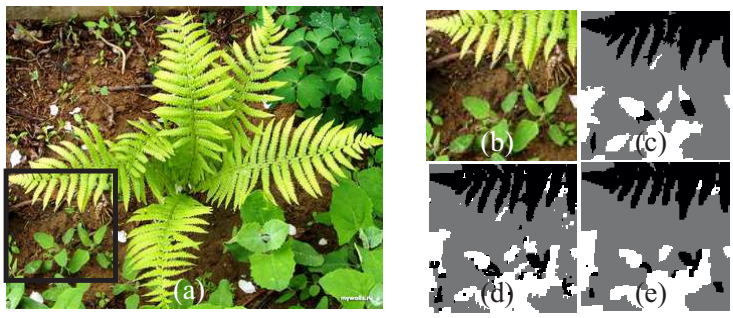


Рисунок 5.5.: (a) – исходное изображение папоротников. (b) – увеличенная часть исходного изображения. (c) – результат алгоритма SMR: найден глобальный минимум энергии; алгоритм CWD может получить такой же результат, если дождаться сходимости. (d), (e) – результаты сегментации этого же изображения без потенциалов высоких порядков с различными значениями весов парных потенциалов, вычисленные при помощи алгоритма α-расширение.

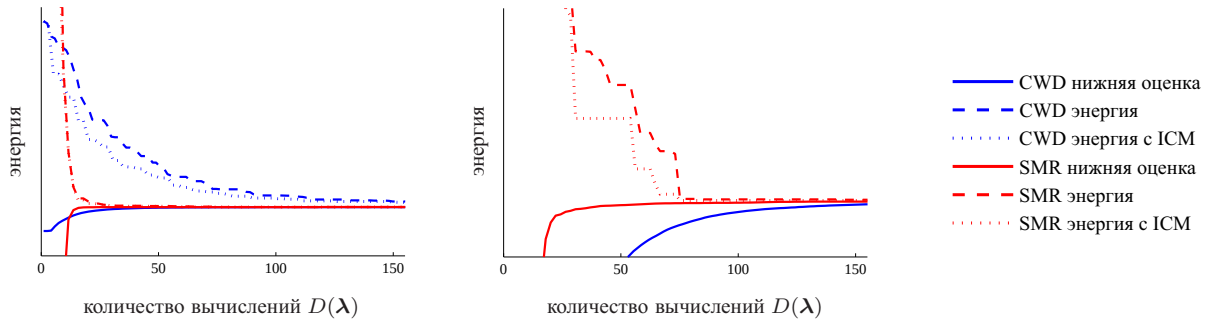


Рисунок 5.6.: Энергии и нижние оценки, полученные при помощи максимизации нижних оценок SMR и CWD, построенных для энергии задачи сегментации изображения папоротников. График (b) является увеличенным фрагментам графика (a). Таблица (c) поясняет значение каждой изображенной кривой.

5.3. Произвольные парно-сепарабельные энергии

Данный раздел посвящён экспериментальному исследованию применимости алгоритма NSMR для случая неассоциативных парно-сепарабельных энергий. Сначала описывается эксперимент по оценке зависимости размера зазора от доли отталкивающих потенциалов. Далее описываются эксперименты, сравнивающие подходы NSMR, DD TRW и TRW-S.

Зазор между прямой и двойственной задачами. Величина зазора между глобальным минимумом задачи минимизации дискретной энергии и максимумом нижней оценки является важным показателем, позволяющим оценить сложность задачи. Если величина зазора равна 0, то обычно методы, основанные на релаксации Лагранжа, позволяют точно решить исходную за-

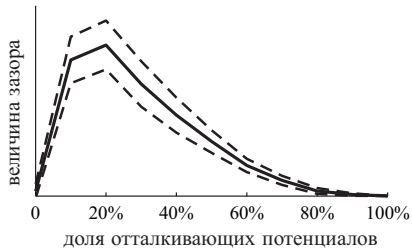


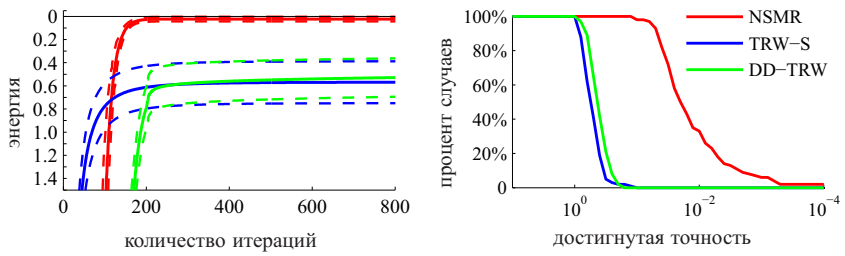
Рисунок 5.7.: Зависимость величины зазора между глобальным минимумом задачи минимизации дискретной энергии и максимумом нижней оценки. Сплошная линия показывает среднее значения, пунктирные линии – одно стандартное отклонение.

дачу. Принято считать, что чем больше зазор, тем «сложнее» конкретная задача минимизации энергии.

Величина зазора между задачами минимизации энергии и двойственной к ней сильно зависит от соотношений между потенциалами. В работе [65] было показано, что если в энергии присутствуют одновременно притягивающие (ассоциативные) и отталкивающие потенциалы, то размер зазора между прямой и двойственной задачами увеличивается при росте относительной силы парных потенциалов. В данном эксперименте проведено исследование влияние другого фактора – доли отталкивающих потенциалов, а именно потенциалов Поттса с положительным коэффициентом: $C_{ij}[x_i = x_j]$, $C_{ij} > 0$.

Рис. 5.7 показывает зависимость оценки величины зазора от доли отталкивающих потенциалов (среди всех парных потенциалов). Оценка величины зазора проводилась при помощи алгоритма TRW-S в авторской реализации [65]. Для каждого значения доли отталкивающих потенциалов генерировались 20 модельных энергий: граф 4-х связная решётка размера 50 на 50, 10 возможных меток каждой переменной, унарные потенциалы генерировались из стандартного нормального распределения $\mathcal{N}(0, 1)$, парные потенциалы – потенциалы Поттса, модули коэффициентов выбирались как модули случайных величины с распределением $\mathcal{N}(0, 2)$, знак коэффициентов Поттса определялся случайных образом так, чтобы контролировалась доля отрицательных коэффициентов. Рис. 5.7 показывает, что величина зазора наибольшая, когда доля отталкивающих потенциалов составляет 20-30%. Можно заметить, что при большом количестве отталкивающих потенциалов величина зазора очень мала.

Сравнение NSMR с DD TRW и TRW-S. В рамках этого эксперимента проводилось сравнение алгоритма NSMR с алгоритмами TRW-S и DD TRW. Набор энергий для сравнения содержал 100 модельных энергий, сгенерированных аналогично предыдущему эксперименту. Доля



(a) (b)

Рисунок 5.8.: Результаты эксперимента по сравнению подхода NSMR с подходом DD TRW и алгоритмом TRW-S. График (a) показывает зависимость нижней оценки глобального минимума энергии, полученной при помощи NSMR, TRW-S и DD TRW от количества итераций. Ось ординат этого графика соответствует величине зазора между текущим значением нижней оценки и точным решением линейной релаксации. Каждая из энергий нормирована так, чтобы разность между значением энергии на разметке, найденной TRW-S, и точным решением линейной релаксации составляла 100, и значение линейной релаксации в оптимуме составляло 0. Ось абсцисс соответствует номеру текущей итерации. Сплошные линии показывают значение, усреднённое по всем энергиям набора; пунктирные линии – одно стандартное отклонение. График (b) показывает долю энергий тестового набора, на которых удалось достичь заданного значения точности за 5000 итераций. Ось абсцисс графика (b) отображается в логарифмической шкале, её единицы измерения соответствуют единицам измерения оси ординат графика (a).

отталкивающих потенциалов Поттса была выставлена на уровне 30%. В качестве TRW-S [65] была использован оригинальная реализация автора. В алгоритме DD TRW в качестве разбиения графа задачи на подграфы было выбрано стандартное разбиение на строки и столбцы. В качестве метода оптимизации нижней оценки в алгоритмах NSMR и DD TRW был выбран метод субградиентного подъёма (4.1) с адаптивным выбором длины шага (4.5) с переключением на неадаптивную схему (4.3) для обеспечения теоретических гарантий сходимости.

Рис. 5.8a показывает зависимость значений нижних оценок всех трёх методов от номера итерации. Рис. 5.8b показывает долю задач, на которых удалось достичь разных уровней точности за 5000 итераций каждого из методов. На графиках видно, что TRW-S находит наилучшие нижние оценки на начальных итерациях, но после этого застревает в покоординатных максимумах, т. е. не сходится к глобальным максимумам нижних оценок. Плохие результаты DD TRW могут быть объяснены тем, что число переменных в нижней оценке DD TRW в 10 раз больше, чем число переменных в нижней оценке NSMR. Тем не менее, возможно, алгоритм DD TRW можно улучшить за счёт использования другой схемы разбиения графа энергии на подграфы-деревья.

5.4. Глобальные ограничения

Этот раздел посвящён описанию экспериментов по использованию подхода SMR с глобальными ограничениями. Сначала проводится сравнение данного подхода с существующими альтернативами, а затем на нескольких реальных задачах демонстрируется применимость метода.

5.4.1. Сравнение с аналогами

Здесь приводится описание эксперимента по сравнению подхода SMR с глобальными ограничениями с существующими альтернативами.

Данные. Набор энергий для этого эксперимента генерируется аналогично наборам для экспериментов, описанных ранее. Набор состоит из 50 энергий; графы всех энергий, задающие парные и унарные потенциалы, являются 4-х связными решётками размера 50 на 50; все переменные могут принимать одно из 10 значений. Унарные потенциалы генерируются как независимые случайные величины со стандартным нормальным распределением: $\theta_{ip} \sim \mathcal{N}(0, 1)$, $i \in \mathcal{V}$, $p \in \mathcal{P}$. Парные потенциалы являются взвешенными потенциалами Поттса, веса которых генерируются как абсолютные значения случайных величин, сгенерированных из нормального распределения $\mathcal{N}(0, 0.5)$. В качестве глобальных ограничений выбраны жёсткие ограничения на количество переменных, принимающих каждое значение $1, \dots, 10$ (2.21). Требуемые количества выбраны таким образом, чтобы существенно отличаться от размеров классов в оптимальном решении без ограничений: для метки $p \in \mathcal{P} = 1, \dots, 10$ требуемое количество выбрано как $|\mathcal{V}|p/\sum_{q=1}^{10} q$.

Альтернативные подходы. Существует немного альтернативных подходов, позволяющих учитывать линейные глобальные ограничения общего вида. В рамках данного эксперимента будут рассмотрены 2 из них:

1. релаксация Лагранжа линейных ограничений как внешний цикл над каким-либо методом, позволяющим находить нижнюю оценку на глобальный минимум энергии.
2. двойственная декомпозиция (релаксация Лагранжа ограничений согласованности подзадач) на задачу минимизации энергии и задачу, отдельно учитывающую глобальные ограничения (аналогично подходу Marginal Probability Field, MPF [135]).

Для реализации метода 1 была выбрана стандартная линейная релаксация, приближённо решаемая при помощи алгоритма TRW-S. Во внешнем цикле при этом решается следующая

оптимизационная задача:

$$\begin{aligned} \max_{\boldsymbol{\xi},\boldsymbol{\pi}} \quad & LB_{TRW}(\tilde{\boldsymbol{\theta}}) - \sum_{m=1}^{M} \xi_m c^m - \sum_{k=1}^{K} \pi_k d^k, \\ \text{s.t.} \quad & \tilde{\theta}_{jp} = \theta_{jp} + \sum_{m=1}^{M} \xi_m w_{jp}^m + \sum_{k=1}^{K} \pi_k v_{jp}^k, \quad \forall j \in \mathcal{V},\ \forall p \in \mathcal{P}, \\ & \tilde{\theta}_{ij,pq} = \theta_{ij,pq}, \quad \forall \{i,j\} \in \mathcal{E},\ \forall p,q \in \mathcal{P}, \\ & \pi_k \geq 0, \quad \forall k = 1,\dots,K. \end{aligned} \tag{5.1}$$

Здесь $LB_{TRW}(\tilde{\boldsymbol{\theta}})$ – стандартная линейная релаксация энергии с потенциалами $\tilde{\boldsymbol{\theta}}$, вычисляемая при помощи TRW-S. Фактически, задача (5.1) решается при помощи метода покоординатного подъёма: чередуется максимизация нижней оценки LB_{TRW} и максимизация целевой функции (5.1) по переменным $\boldsymbol{\xi}$ и $\boldsymbol{\pi}$. Далее данный метод называется GTRW.

Метод 2 фактически является упрощённой версией подхода Marginal Probability Field (MPF) [135]. Пусть $LB_{TRW}(\hat{\boldsymbol{\theta}}(\boldsymbol{\mu}))$ – решение стандартной линейной релаксации задачи минимизации энергии без глобальных ограничений, где унарные потенциалы модифицированы: $\hat{\theta}_{jp}(\boldsymbol{\mu})) = \theta_{jp} - \mu_{jp}$, $j \in \mathcal{V}$, $p \in \mathcal{P}$, а парные потенциалы не изменены: $\hat{\theta}_{ij,pq}(\boldsymbol{\mu})) = \theta_{ij,pq}$, $\{i,j\} \in \mathcal{E}$, $p,q \in \mathcal{P}$. Пусть

$$\begin{aligned} \Phi_G(\boldsymbol{\mu}) = \min_{\boldsymbol{y}} \quad & \langle \boldsymbol{\mu}, \boldsymbol{y} \rangle, \\ \text{s.t.} \quad & \sum_{j \in \mathcal{V}} y_{jp} = c_p \geq 0, \quad \forall p \in \mathcal{P}, \\ & 0 \leqslant y_{jp} \leqslant 1, \quad \forall j \in \mathcal{V},\ \forall p \in \mathcal{P}. \end{aligned} \tag{5.2}$$

Задача (5.2) является транспортной задачей (транспортировка из $|\mathcal{V}|$ вершин в $|\mathcal{P}|$ вершин) и имеет решение, только если ограничения сбалансированы: $\sum_{p \in \mathcal{P}} c_p = |\mathcal{V}|$.

Сумма $LB_{TRW}(\hat{\boldsymbol{\theta}}(\boldsymbol{\mu})) + \Phi_G(\boldsymbol{\mu})$ является вогнутой кусочно-линейной функцией по переменным $\boldsymbol{\mu}$ и может быть максимизирована методами выпуклой оптимизации. В рамках данного эксперимента функция $LB_{TRW}(\hat{\boldsymbol{\theta}}(\boldsymbol{\mu}))$ вычислялась при помощи алгоритма TRW-S, а функция $\Phi_G(\boldsymbol{\mu})$ при помощи симплекс-метода решения задач линейного программирования. Далее данный метод называется MPF.

Результаты. Рис. 5.9 показывает графики сходимости нижних оценок, посчитанных при помощи алгоритмов SMD, GTRW и MPF, усреднённые по сгенерированному набору энергий. Для алгоритма MPF время работы алгоритма решения транспортной задачи не учитывается (т. к. данное время очень сильно зависит от конкретной реализации). График 5.9а показывает нижние

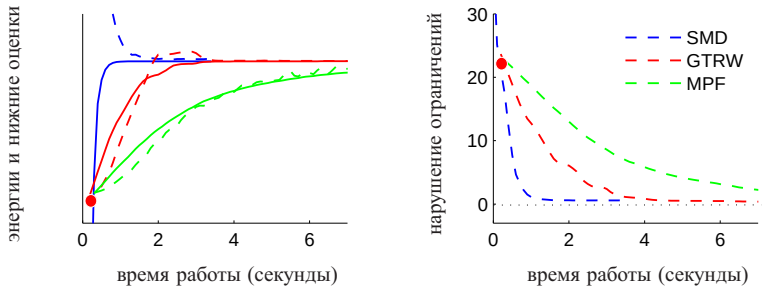


Рисунок 5.9.: Сравнение алгоритмов SMD, GTRW и MPF на наборе синтетических данных. График (a): сплошные линии соответствуют нижним оценкам, пунктирные линии соответствуют энергиям разметок (которые, вообще говоря, не удовлетворяют глобальным ограничениям), полученным на каждой итерации; график (b) показывает текущее общее нарушение ограничений (количество переменных, значения которых надо изменить, чтобы удовлетворить ограничениям). На обоих графиках по оси абсцисс отложено время работы в секундах. Синие линии везде соответствуют SMD, красные – GTRW, зелёные – MPF. Красные круги на обоих графиках показывают результата алгоритма TRW-S, как «начальную точку» GTRW и MPF.

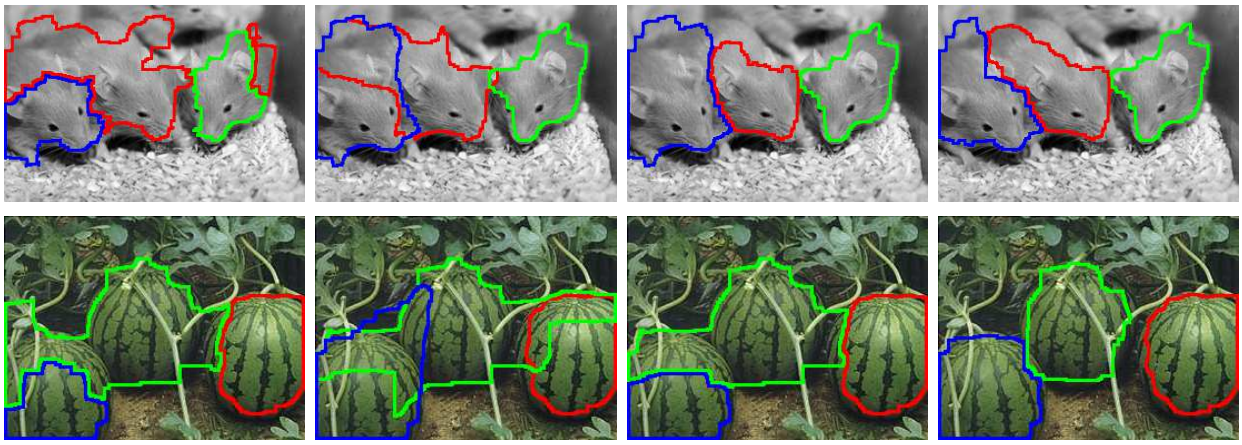
оценки и прямые решения[7] в зависимости от времени работы для всех трёх методов. Заметим, что здесь энергия может быть меньше нижней оценки, поскольку разметка, на которой она вычисляется, вообще говоря, не удовлетворяет глобальным ограничениям. Более того, а алгоритмах GTRW и MPF энергия текущего решения увеличивается, поскольку оба метода начинают с нижней оценки, полученной TRW-S без учёта ограничений, и далее пытается согласовать решение с этими ограничениями. График 5.9b показывает общее нарушение ограничений текущего решения. Исходя из рис. 5.9, можно сделать вывод, что все методы сходятся к одной точке, но SMD делает это намного быстрее.

5.4.2. Применимость метода на реальных данных

В данном разделе показывается способность подхода SMR учитывать глобальные линейные ограничения в некоторых реальных задачах.

Сегментация изображений. Подход SMD позволяет учитывать не только глобальные линейные ограничения, но и любые ограничения, которые могут наложены путём таких модификации

[7] Выбор прямого решения осуществлялся при помощи следующих эвристик: в алгоритме SMD всем пикселям с конфликтующими метками в одной компоненте связности присваивалась одна случайная метка из множества конфликтующих меток; в алгоритмах GTRW и MPF в качестве текущей разметки выбиралась разметка, найденная в подзадаче TRW-S.



(a) (b) (c) (d)

Рисунок 5.10.: Сегментация изображений с учётом глобальных ограничений. Столбец (a) – результаты алгоритмов TRW-S и α-расширение без глобальных ограничений (не отличимы на глаз). Столбец (b) – результаты независимой сегментации каждого из объектов с ограничением на «звёздность» формы. Столбец (c) – результаты алгоритма SMD и α-расширение, полученные с учётом ограничений на «звёздность» формы. Столбец (d) – результаты алгоритма SMD, учитывающего как ограничения на «звёздность формы», так и ограничения на равенство площадей меток, ассоциированных с объектами, друг другу.

энергии, что она остаётся субмодулярной. Рассмотрим возможность совмещения ограничений разных типов на примере задачи сегментации изображений. Примером ограничений, которые можно учесть при помощи модификация субмодулярных энергий служат ограничения «звёздности» объектов (star-shape prior) [125]. Эксперимент проводится с использованием семян объектов – для каждого сегмента выделяется небольшая область, которая гарантированно относится к данному сегменту (аналогично работе [25]). Выделенные области используются для сбора статистики о цвете пикселей сегментов, а также для задания центров «звёздности».

В качестве унарных потенциалов, как обычно, используется отрицательный логарифм правдоподобия цветовой модели, построенной по семенам (смеси из 5 нормальных распределений). В качестве парных потенциалов используются взвешенные потенциалы Поттса: $C_{ij} = 2 + 20 \exp\left(-\frac{\|I_i - I_j\|_2^2}{2\sigma_{ij}^2}\right)$, $\{i, j\} \in \mathcal{E}$. Здесь I_i – цвет пикселя $i \in \mathcal{V}$. Параметр σ_{ij} выставляется для каждого ребра, как средний модуль разницы цветов соседних пикселей в окрестности 20 на 20.

Ограничение «звёздности» – это ограничение на форму объекта следующего вида: на любом луче, исходящем из центра «звёздности» $c \in \mathcal{V}$, метка объекта 1 не может стоять после метки фона 0. Данное ограничение может быть записано чрез парные потенциалы следующим

образом:

$$S_{ij} = \begin{cases} 0, & \text{если } f_i = f_j, \\ \infty, & \text{если } f_i = 1 \text{ и } f_j = 0, \\ 0, & \text{если } f_i = 0 \text{ и } f_j = 1, \end{cases}$$

где $j \in \mathcal{V}$ лежит на луче из центра c в вершину i. Подробно ограничение «звёздности» описано в работе [125].

Рис. 5.10 содержит качественные результаты алгоритмов сегментации изображений, работающих с и без учёта глобальных ограничений («звёздность» и равенство площадей). Результаты без учёта глобальных ограничений (рис. 5.10a) получены алгоритмом α-расширение (не отличимы на глаз от результатов алгоритмов TRW-S и SMD). Результаты на рис. 5.10b построены при помощи независимой сегментации каждого из объектов при помощи алгоритма из работы [125], заключающегося в построении минимального разреза специального графа. Рис. 5.10c содержит сегментации, построенные при помощи алгоритмов SMD и α-расширение (сегментации не отличимы), учитывающие ограничение «звёздности» для всех классов, за исключением фона. На рис. 5.10d приведены результаты алгоритма SMD, учитывающего одновременно ограничения на «звёздность» каждого из объектов и ограничения на равенство площадей объектов.

Сегментация магнитограмм. Данный эксперимент показывает применимость метода SMD для учёта линейных ограничений общего вида: равенство потоков (2.25). Примером приложения, где естественным образом возникают ограничения на равенство потоков является задача сегментации магнитограмм Солнца. Рис. 5.11a,b содержат изображения солнца в ультрафиолетовом и магнитном спектрах. Изображение 5.11b также называется магнитограммой. В областях повышенной солнечной активности (солнечные пятна) (рис. 5.11c) интенсивность магнитного поля существенно выше (участки с положительным направлением силовых линий магнитного поля отмечены белым, с отрицательным – чёрным). Чем больше отклонение от среднего уровня, тем сильнее отличается цвет соответствующего пикселя от серого. Сегментация солнечной поверхности на спокойные области, области поля с положительным и отрицательным направлением силовых линий, а также анализ их взаимного расположения позволяет получать важные признаки, на основе которых можно осуществлять прогноз солнечных вспышек. Известно, что в некоторой окрестности солнечного пятна суммарный положительный поток магнитного поля примерно равен суммарному отрицательному потоку. Ограничение равенства потоков (2.25) является частным случаем линейных ограничений и может быть учтено при помощи подхода SMD.

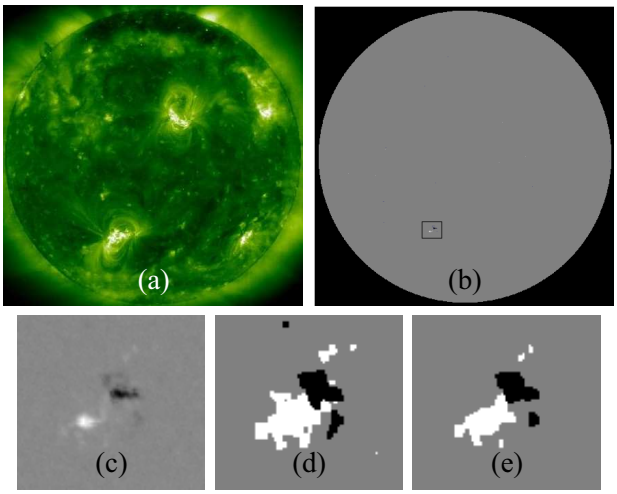


Рисунок 5.11.: Задача сегментации магнитограмм Солнца. (a) – изображение Солнца в ультрафиолетовом диапазоне; (b) – магнитограмма Солнца: белые точки соответствуют областям с положительным направлением силовых линий магнитного поля, чёрные точки – областям с отрицательным направлением силовых линий, серые пиксели – областям с относительно малой активностью. (c) – увеличенная область с повышенной активностью (солнечное пятно). (d) – результат сегментации области (c), полученный при помощи алгоритма α-расширения, не учитывающего ограничения на равенство потоков: положительный поток = 83794, отрицательный поток = -71021, разница = 12773; (e) – результат алгоритма SMD, учитывающего ограничения на равенство потоков: положительный поток = 70972, отрицательный поток = -67341, разница = 3630.

Рис. 5.11d–e показывают результаты сегментации региона солнечных пятен на три области без учёта ограничений равенство потока и с их учётом.

Сегментация изображений с неассоциативными парными потенциалами и глобальными ограничениями. Неассоциативные потенциалы для сегментации изображений используются редко, поскольку приводят к оптимизационным задачам, которые сложно решать. Для демонстрации эффекта глобальных ограничений в такой ситуации построим искусственное изображение с семенами 5.12a, которое требуется отсегментировать на символы и фон. Все объекты и фон обладают одинаковыми распределениями цветов, поэтому унарные потенциалы не дают никакой информации (за исключением учёта семян). В качестве парных потенциалов на рёбрах со слабыми перепадами цвета используются притягивающие потенциалы Поттса $C_{ij} = -1$, на рёбрах же с сильным контрастом используются отталкивающие потенциалы Поттса $C_{ij} = 1$. Сильный контраст отделяется от слабого при помощи отсечения модуля разности интенсивности по порогу 0.1 (чёрные пиксели обладают интенсивностью 0, белые – 1):

$$\theta_{ij}(x_i, x_j) = -[|I_i - I_j| < 0.1][x_i = x_j] + [|I_i - I_j| \geqslant 0.1][x_i = x_j], \quad \{i, j\} \in \mathcal{E},$$

(a) (b) (c) (d)

Рисунок 5.12.: Сегментация искусственного изображения по заданным пользователем семенам. (a) – исходное изображение с семенами; (b) – результат алгоритма TRW-S; (c) – результат алгоритма NSMR; (d) – результат алгоритма NSMR с глобальными ограничениями на площади объектов.

где I_i – интенсивность пикселя $i \in \mathcal{V}$. Такая энергия является сложной, поскольку унарные потенциалы очень слабые, и присутствуют одновременно притягивающие и отталкивающие парные потенциалы. Рис. 5.12b-c показывают результаты алгоритмов TRW-S и NSMR на данной задаче (фрагментированная сегментация). Рис. 5.12d показывает результаты алгоритма NSMR с глобальными линейными ограничениями, задающими равенство площадей объектов правильным значениям (2.21).

Заключение

Основные результаты данной работы заключаются в следующем:

1. Разработан новый подход к решению задачи минимизации энергий: субмодулярная релаксация (SMR). Подход SMR основан на релаксации Лагранжа ограничений целостности. Основное отличие подхода SMR от существующих методов, основанных на релаксации Лагранжа и двойственной декомпозиции (DD TRW [73], CWD [71]), состоит в том что не происходит разбиение графа задачи на подграфы.

2. Проведено теоретическое исследование подхода SMR. Приведена точная формулировка линейной релаксации, значение решения которой равно максимуму нижней оценки SMR. Доказано, что в случае парно-сепарабельных энергий с ассоциативными парными потенциалами построенная линейная релаксация эквивалентна стандартной (решаемой DD TRW). Также показано, что в случае перестановочных потенциалов Поттса высоких порядков построенная линейная релаксация эквивалентна релаксации, решаемой методом CWD.

3. Построено обобщение подхода SMR на случай несубмодулярного лагранжиана: разработан подход NSMR, сформулирована задача линейного программирования, эквивалентная ему.

4. Исследован вопрос применимости различных методов выпуклой оптимизации для решения задачи поиска наилучшей нижней оценки на глобальный минимум энергии из семейства оценок, построенных подходом SMR. Разработан алгоритм покоординатного подъёма, обеспечивающий монотонное улучшение оценки в случае парно-сепарабельных ассоциативных потенциалов.

5. Проведено экспериментальное исследование подхода SMR, включающее в себя его сравнение с аналогами: DD TRW [73], CWD [71], TRW-S [65]. Показано, что в ряде случаев алгоритм SMR превосходит аналоги по скорости работы. Также показана возможность учёта некоторых видов глобальных ограничений в рамках подхода SMR.

Список рисунков

Список таблиц

Литература

1. Васильев Ф. П. Методы оптимизации. Москва: Факториал Пресс, 2002.

2. Двоенко С. Д. Методы распознавания образов в массивах взаимосвязанных данных. Дисс. докт. физ.-мат. наук. 2001.

3. Двоенко С. Д., Копылов А. В., Моттль В. В. Задача распознавания образов в массивах взаимосвязанных объектов. Постановка задачи распознавания и основные предположения // Автоматика и телемеханика. 2004. № 1. С. 143–158.

4. Двоенко С. Д., Копылов А. В., Моттль В. В. Задача распознавания образов в массивах взаимосвязанных объектов. Алгоритм распознавания // Автоматика и телемеханика. 2005. № 12. С. 162–176.

5. Коваль В. К., Шлезингер М. И. Двумерное программирование в задачах анализа изображений // Автоматика и телемеханика. 1976. Т. 8. С. 149–168.

6. Кормен Т. Х., Лейзерсон Ч. И., Ривест Р. Л., Штайн К. Алгоритмы: построение и анализ. 3-е изд. Моксва: «Вильямс», 2013.

7. Осокин А. А., Ветров Д. П. Решение задач оптимизации на марковских случайных полях с помощью разложения, сохраняющего структуру графа // Доклады 15-ой Всероссийской конференции «Математические методы распознавания образов». 2011. С. 207–210.

8. Шлезингер М. И. Синтаксический анализ двумерных зрительных сигналов в условиях помех // Кибернетика. 1976. Т. 4. С. 113–130.

9. Шлезингер М. И., Гигиняк В. В. Решение (MAX,+)-задач структурного распознавания с помощью их эквивалентных преобразований // Управляющие системы и машины. 2007. № 1,2.

10. Шор Н. З. Методы минимизации недифференцируемых функций и их приложения. Киев: Наукова думка, 1979.

11. Alahari K., Kohli P., Torr P. H. S. Dynamic Hybrid Algorithms for MAP Inference in Discrete MRFs // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2010. Vol. 32, no. 10. P. 1846–1857.

12. Andres B., Köthe U., Kroeger T., Helmstaedter M., Briggman K. L., Denk W., Hamprecht F. A. 3D segmentation of SBFSEM images of neuropil by a graphical model over supervoxel boundaries // Medical Image Analysis. 2012. Vol. 16, no. 4. P. 796–805.

13. Anstreicher K. M., Wolsey L. A. Two "well-known" properties of subgradient optimization // Mathematical Programming. 2009. Vol. 120. P. 213–220.

14. Arora C., Banerjee S., Kalra P., Maheshwari S. N. Generic Cuts: An Efficient Algorithm for Optimal Inference in Higher Order MRF-MAP // European Conference on Computer Vision (ECCV). 2012.

15. Barahona F., Anbil R. The volume algorithm: producing primal solutions with a subgradient method // Mathematical Programming. 2000. Vol. 87, no. 3. P. 385–399.

16. Batra D., Gallagher A., Parikh D., Chen T. Beyond Trees: MRF Inference via Outer-Planar Decomposition // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010.

17. Batra D., Nowozin S., Kohli P. Tighter Relaxations for MAP-MRF Inference: A Local Primal-Dual Gap based Separation Algorithm // International Conference on Artificial Intelligence and Statistics (AISTATS). 2011.

18. Bertsekas D. P., Nedic A., Ozdaglar A. E. Convex Analysis and Optimization. Athena Scientific, 2003.

19. Besag J. E. On the Statistical Analysis of Dirty Pictures // Journal of the Royal Statistical Society, Series B. 1986. Vol. 48, no. 3. P. 259–302.

20. Bishop C. M. Pattern recognition and machine learning. Springer, 2006.

21. Blake A., Zisserman A. Visual Reconstruction. MIT Press, 1987.

22. Boros E., Hammer P. L. Pseudo-boolean optimization // Discrete Applied Mathematics. 2002. Vol. 123. P. 155–225.

23. Boykov Y., Kolmogorov V. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2004. Vol. 26, no. 9. P. 1124–1137.

24. Boykov Y., Veksler O., Zabih R. Fast approximate energy minimization via graph cuts // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2001. Vol. 23, no. 11. P. 1222–1239.

25. Boykov Y., Funka-Lea G. Graph Cuts and Efficient N-D Image Segmentation // International Journal of Computer Vision (IJCV). 2006. Vol. 70, no. 2. P. 109–131.

26. Boykov Y., Jolly M.-P. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images // International Conference on Computer Vision (ICCV). 2001.

27. Chekuri C., Khanna S., Naor J., Zosin L. A linear programming formulation and approximation algorithms for the metric labeling problem // SIAM Journal on Discrete Mathematics. 2004. Vol. 18, no. 3. P. 608–625.

28. Chen Y., Ye X. Projection Onto A Simplex: Tech. Rep.: 1101.6081: arXiv, 2011.

29. Christoudias C., Georgescu B., Meer P. Synergism in low-level vision // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2002.

30. Darbon J. Global optimization for first order Markov random fields with submodular priors // Discrete Applied Mathematics. 2009. Vol. 157, no. 16. P. 3412–3423.

31. Delong A., Osokin A., Isack H. N., Boykov Y. Fast Approximate Energy Minimization with Label Costs // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010.

32. Delong A., Osokin A., Isack H. N., Boykov Y. Fast Approximate Energy Minimization with Label Costs // International Journal of Computer Vision (IJCV). 2012. Vol. 96, no. 1. P. 1–27.

33. Delong A., Veksler O., Osokin A., Boykov Y. Minimizing Sparse High-Order Energies by Submodular Vertex-Cover // Advances in Neural Information Processing Systems (NIPS). 2012.

34. Duchi J., Tarlow D., Elidan G., Koller D. Using Combinatorial Optimization within Max-Product Belief Propagation // Advances in Neural Information Processing Systems (NIPS). 2006.

35. Felzenszwalb P., Huttenlocher D. Efficient Belief Propagation for Early Vision // International Journal of Computer Vision (IJCV). 2006. Vol. 70, no. 1. P. 41–54.

36. Felzenszwalb P., Huttenlocher D. Distance Transforms of Sampled Functions // Theory of Computing. 2012. Vol. 8, no. 19. P. 415–428.

37. Felzenszwalb P. F., Girshick R. B., McAllester D., Ramanan D. Object Detection with Discriminatively Trained Part Based Models // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2010. Vol. 32, no. 9. P. 1627–1645.

38. Freedman D., Drineas P. Energy minimization via graph cuts: settling what is possible // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005.

39. Freedman D., Zhang T. Interactive Graph Cut Based Segmentation With Shape Priors // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005.

40. Frey B. J., Dueck D. Clustering by Passing Messages Between Data Points // Science. 2007. Vol. 315. P. 972–976.

41. Geman S., Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 1984. Vol. 6. P. 721–741.

42. Globerson A., Jaakkola T. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations // Advances in Neural Information Processing Systems (NIPS). 2007.

43. Goldberg A. V., Hed S., Kaplan H., Tarjan R. E., Werneck R. F. Maximum Flows By Incremental Breadth-First Search // European Symposium on Algorithms (ESA). 2011. P. 457–468.

44. Greig D., Porteous B., Seheult A. Exact maximum a posteriori estimation for binary images // Journal of the Royal Statistical Society, Series B. 1989. Vol. 51, no. 2. P. 271–279.

45. Haarala N., Miettinen K., Mäkelä M. M. Globally Convergent Limited Memory Bundle Method for Large-Scale Nonsmooth Optimization // Mathematical Programming. 2007. Vol. 109, no. 1. P. 181–205.

46. Hammer P. L. Some network flow problems solved with pseudo-Boolean programming // Operations Research. 1965. Vol. 13. P. 388–399.

47. Hammer P. L., Hansen P., Simeone B. Roof duality, complementation and persistency in quadratic 0–1 optimization // Mathematical Programming. 1984. Vol. 28, no. 2. P. 121–155.

48. Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks // Science. 2006. Vol. 313, no. 5786. P. 504–507.

49. Ishikawa H. Transformation of General Binary MRF Minimization to the First Order Case // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2011. Vol. 33, no. 6. P. 1234–1249.

50. Ishikawa H. Exact Optimization for Markov Random Fields with Convex Priors // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2003. Vol. 25, no. 10. P. 1333–1336.

51. Iwata S., Orlin J. B. A simple combinatorial algorithm for submodular function minimization // ACM-SIAM Symposium on Discrete Algorithms (SODA). 2009.

52. Kappes J., Savchynskyy B., Schnörr C. A Bundle Approach To Efficient MAP-Inference by Lagrangian Relaxation // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012.

53. Kappes J. H., Schmidt S., Schnörr C. MRF Inference by k-Fan Decomposition and Tight Lagrangian Relaxation // European Conference on Computer Vision (ECCV). 2010.

54. Kappes J., Andres B., Hamprecht F., Schnörr C., Nowozin S., Batra D., Kim S., Kausler B., Lellmann J., Komodakis N., Rother C. A Comparative Study of Modern Inference Techniques for Discrete Energy Minimization Problems // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013.

55. Kappes J., Speth M., Reinelt G., Schnörr C. Towards Efficient and Exact MAP-Inference for Large Scale Discrete Computer Vision Problems via Combinatorial Optimization // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013.

56. Kiwiel K. An aggregate subgradient method for nonsmooth convex minimization // Mathematical Programming. 1983. Vol. 27. P. 320–341.

57. Kiwiel K. C. Proximity control in bundle methods for convex nondifferentiable minimization // Mathematical Programming. 1990. Vol. 46, no. 1-3. P. 105–122.

58. Kleinberg J. M., Tardos É. Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields // Journal of the ACM (JACM). 2002. Vol. 49, no. 5. P. 616–639.

59. Kohli P., Torr P. Dynamic graph cuts for efficient inference in markov random fields // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2007. Vol. 29, no. 12. P. 2079–2088.

60. Kohli P., Kumar M. P., Torr P. $\mathcal{P}^3$&Beyond: Move Making Algorithms for Solving Higher Order Functions // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2008. Vol. 31, no. 9. P. 1645–1656.

61. Kohli P., Ladickỳ L., Torr P. Robust higher order potentials for enforcing label consistency // International Journal of Computer Vision. 2009. Vol. 82, no. 3. P. 302–324.

62. Kohli P., Torr P. H. S. Measuring uncertainty in graph cut solutions // Computer Vision and Image Understanding. 2008. Vol. 112, no. 1. P. 30–38.

63. Kohli P., Shekhovtsov A., Rother C., Kolmogorov V., Torr P. On Partial Optimality in Multi-label MRFs // International Conference on Machine Learning (ICML). 2008.

64. Kohli P., Osokin A., Jegelka S. A Principled Deep Random Field Model for Image Segmentation // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013.

65. Kolmogorov V. Convergent Tree-reweighted Message Passing for Energy Minimization // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2006. Vol. 28, no. 10. P. 1568–1583.

66. Kolmogorov V., Rother C. Minimizing non-submodular functions with graph cuts – a review // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2007. Vol. 29, no. 7. P. 1274–1279.

67. Kolmogorov V., Wainwright M. On the Optimality of Tree-reweighted Max-product Message Passing // Uncertainty in Artificial Intelligence (UAI). 2005.

68. Kolmogorov V., Zabih R. What energy functions can be minimized via graph cuts? // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2004. Vol. 26, no. 2. P. 147–159.

69. Kolmogorov V., Boykov Y., Rother C. Applications of parametric maxflow in computer vision // International Conference on Computer Vision (ICCV). 2007.

70. Komodakis N., Paragios N. Beyond Loose LP-relaxations: Optimizing MRFs by Repairing Cycles // European Conference on Computer Vision (ECCV). 2008.

71. Komodakis N., Paragios N. Beyond Pairwise Energies: Efficient Optimization for Higher-Order MRFs // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009.

72. Komodakis N., Tziritas G., Paragios N. Fast, Approximately Optimal Solutions for Single and Dynamic MRFs // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2007.

73. Komodakis N., Paragios N., Tziritas G. MRF Energy Minimization and Beyond via Dual Decomposition. // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2011. Vol. 33, no. 3. P. 531–552.

74. Kovtun I. Partial Optimal Labeling Search for a NP-Hard Subclass of (max,+) Problems // DAGM Symposium. 2003.

75. Krizhevsky A., Sutskever I., Hinton G. E. ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems (NIPS). 2012.

76. Kropotov D., Laptev D., Osokin A., Vetrov D. Variational segmentation algorithms with label frequency constraints // Pattern Recognition and Image Analysis. 2010. Vol. 20. P. 324–334.

77. Kropotov D., Laptev D., Osokin A., Vetrov D. Signal Segmentation with Label Frequency Constraints using Dual Decomposition Approach for Hidden Markov Models // Intellectualization of information processing (IIP). 2010. P. 403–406.

78. Kumar M. P., Torr P. H. S., Zisserman A. Solving Markov random fields using second order cone programming relaxations // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2006.

79. Kumar M. P., Kolmogorov V., Torr P. H. S. An Analysis of Convex Relaxations for MAP Estimation of Discrete MRFs // Journal of Machine Learning Research (JMLR). 2009. Vol. 10. P. 71–106.

80. Ladický L., Russel C., Kohli P., Torr P. H. S. Graph Cut based Inference with Co-occurrence Statistics // European Conference on Computer Vision (ECCV). 2010.

81. Lafferty J. D., McCallum A., Pereira F. C. N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data // International Conference on Machine Learning (ICML). 2001.

82. Lauritzen S. L. Graphical Models. Oxford University Press, 1996.

83. LeCun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition // Proceedings of the IEEE. 1998. Vol. 86, no. 11. P. 2278–2324.

84. Lemaréchal C. Lagrangian Relaxation // Computational Combinatorial Optimization / Ed. by M. Jünger, D. Naddef. Vol. 2241 of *Lecture Notes in Computer Science*. 2001. P. 112–156.

85. Lempitsky V., Boykov Y. Global Optimization for Shape Fitting // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2007.

86. Lempitsky V., Kohli P., Rother C., Sharp T. Image Segmentation with A Bounding Box Prior // International Conference on Computer Vision (ICCV). 2009.

87. Lempitsky V., Rother C., Roth S., Blake A. Fusion Moves for Markov Random Field Optimization // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2010. Vol. 32, no. 8. P. 1392–1405.

88. Lempitsky V., Vedaldi A., Zisserman A. A Pylon Model for Semantic Segmentation // Advances in Neural Information Processing Systems (NIPS). 2011.

89. Lempitsky V., Blake A., Rother C. Branch-and-Mincut: Global Optimization for Image Segmentation with High-Level Priors // Journal of Mathematical Imaging and Vision. 2012. Vol. 44, no. 3. P. 315–329.

90. Lewis A. S., Overton M. L. Nonsmooth optimization via quasi-Newton methods // Mathematical Programming. 2013. Vol. 141, no. 1-2. P. 135–163.

91. Lim Y., Jung K., Kohli P. Energy Minimization Under Constraints on Label Counts // European Conference on Computer Vision (ECCV). 2010.

92. MacKay D. J. C. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.

93. Nowozin S., Lampert C. H. Global Interactions in Random Field Models: A Potential Function Ensuring Connectedness // SIAM Journal on Imaging Sciences (SIIMS). 2010. Vol. 3, no. 4.

94. Nowozin S., Rother C., Bagon S., Sharp T., Yao B., Kohli P. Decision tree fields // International Conference on Computer Vision (ICCV). 2009.

95. Nowozin S., Lampert C. Structured Learning and Prediction in Computer Vision. Foundations and Trends in Computer Graphics and Vision no. 3–4. Now publishers, 2011.

96. Osokin A., Vetrov D. Submodular Relaxation for MRFs with High-Order Potentials // Computer Vision – ECCV 2012. Workshops and Demonstrations / Ed. by A. Fusiello, V. Murino, R. Cucchiara. Vol. 7585 of *Lecture Notes in Computer Science*. 2012. P. 305–314.

97. Osokin A., Vetrov D., Kolmogorov V. Submodular Decomposition Framework for Inference in Associative Markov Networks with Global Constraints // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011.

98. Osokin A., Vetrov D., Kolmogorov V. Submodular Decomposition Framework for Inference in Associative Markov Networks with Global Constraints: Tech. Rep.: 1103.1077: arXiv, 2011.

99. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco: Morgan Kaufman, 1988.

100. Pletscher P., Nowozin S., Kohli P., Rother C. Putting MAP back on the map // DAGM Symposium. 2011.

101. Ravikumar P., Lafferty J. Quadratic Programming Relaxations for Metric Labeling and Markov Random Fields MAP Estimation // International Conference on Machine Learning (ICML). 2006.

102. Ravikumar P., Agarwal A., Wainwright M. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes // Journal of Machine Learning Research. 2010. Vol. 11. P. 1043–1080.

103. Rother C., Kolmogorov V., Lempitsky V., Szummer M. Optimizing binary MRFs via extended roof duality // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2007.

104. Rother C., Kohli P., Feng W., Jia J. Minimizing Sparse Higher Order Energy Functions of Discrete Variables // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009.

105. Savchynskyy B., Schmidt S. Getting Feasible Variable Estimates From Infeasible Ones: MRF Local Polytope Study: Tech. Rep.: 1210.4081: arXiv, 2012.

106. Savchynskyy B., Kappes J. H., Schmidt S., Schnörr C. A Study of Nesterov's Scheme for Lagrangian Decomposition and MAP Labeling // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011.

107. Savchynskyy B., Kappes J., Swoboda P., Schnörr C. Global MAP-Optimality by Shrinking the Combinatorial Search Area with Convex Relaxation // Advances in Neural Information Processing Systems (NIPS). 2013.

108. Schmidt M., Alahari K. Generalized Fast Approximate Energy Minimization via Graph Cuts: Alpha-Expansion Beta-Shrink Moves // Uncertainty in Artificial Intelligence (UAI). 2011.

109. Schraudolph N. N., Kamenetsky D. Efficient Exact Inference in Planar Ising Models // Advances in Neural Information Processing Systems (NIPS). 2009.

110. Shalev-Shwartz S., Singer Y., Srebro N., Cotter A. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM // Mathematical Programming. 2011. Vol. 127, no. 1. P. 3–30.

111. Shekhovtsov A., Kohli P., Rother C. Curvature Prior for MRF-based Segmentation and Shape Inpainting // Pattern Recognition / Ed. by A. Pinz, T. Pock, H. Bischof, F. Leberl. Vol. 7476 of *Lecture Notes in Computer Science*. 2012. P. 41–51.

112. Sherali H., Choib G. Recovery of primal solutions when using subgradient optimization methods to solve Lagrangian duals of linear programs // Operations Research Letters. 1996. Vol. 19, no. 3. P. 105–113.

113. Shimony S. E. Finding MAPs for belief networks is NP-hard // Artificial Intelligence. 1994. Vol. 68, no. 2. P. 399–410.

114. Shotton J., Winn J., Rother C., Criminisi A. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation // European Conference on Computer Vision (ECCV). 2006. P. 1–14.

115. Sontag D., Jaakkola T. New Outer Bounds on the Marginal Polytope // Advances in Neural Information Processing Systems (NIPS). 2007.

116. Sontag D., Meltzer T., Globerson A., Weiss Y., Jaakkola T. Tightening LP Relaxations for MAP using Message Passing // Uncertainty in Artificial Intelligence (UAI). 2008.

117. Sontag D., Globerson A., Jaakkola T. Introduction to Dual Decomposition for Inference // Optimization for Machine Learning / Ed. by S. Sra, S. Nowozin, S. J. Wright. MIT Press, 2011.

118. Sun M., Telaprolu M., Lee H., Savarese S. Efficient and Exact MAP Inference using Branch and Bound // International Conference on Artificial Intelligence and Statistics (AISTATS). 2012.

119. Swoboda P., Savchynskyy B., Kappes J., Schnörr C. Partial optimality via iterative pruning for the Potts model // International Conference on Scale Space and Variational Methods in Computer Vision (SSVM). 2013. P. 477–488.

120. Szeliski R., Zabih R., Scharstein D., Veksler O., Kolmogorov V., Agarwala A., Tappen M., Rother C. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2008. Vol. 30, no. 6. P. 1068–1080.

121. Tarlow D., Givoni I., Zemel R. HOP-MAP: Efficient Message Passing with High Order Potentials // International Conference on Artificial Intelligence and Statistics (AISTATS). 2010.

122. Taskar B., Guestrin C., Koller D. Max-Margin Markov Networks // Advances in Neural Information Processing Systems (NIPS). 2003.

123. Taskar B., Chatalbashev V., Koller D. Learning associative Markov networks // International Conference on Machine Learning (ICML). 2004.

124. Tsochantaridis I., Joachims T., Hofmann T., Altun Y. Large Margin Methods for Structured and Interdependent Output Variables // Journal of Machine Learning Research (JMLR). 2005. Vol. 6, no. 9. P. 1453–1484.

125. Veksler O. Star Shape Prior for Graph-Cut Image Segmentation // European Conference on Computer Vision (ECCV). 2008.

126. Veksler O. Multi-label Moves for MRFs with Truncated Convex Priors // International Journal of Computer Vision (IJCV). 2012. Vol. 98, no. 1. P. 1–14.

127. Vetrov D., Osokin A. Graph Preserving Label Decomposition in Discrete MRFs with Selfish Potentials // NIPS Workshop on Discrete Optimization in Machine learning (DISCML NIPS). 2011.

128. Vetrov D., Osokin A. Submodular Decomposition Approach for Inference in Markov Random Fields // Intellectualization of information processing (IIP). 2010. P. 5–8.

129. Wainwright M. J., Jordan M. I. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, no. 1–2. Now publishers, 2008.

130. Wainwright M. J., Jaakkola T. S., Willsky A. S. MAP estimation via agreement on trees: message-passing and linear programming // IEEE Transactions on Information Theory. 2005. Vol. 51, no. 11. P. 3697–3717.

131. Wang P., Shen C., van den Hengel A. A Fast Semidefinite Approach to Solving Binary Quadratic Problems // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013.

132. Werner T. A Linear Programming Approach to Max-sum Problem: A Review // IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). 2007. Vol. 29, no. 7. P. 1165–1179.

133. Werner T. High-arity Interactions, Polyhedral Relaxations, and Cutting Plane Algorithm for Soft Constraint Optimisation (MAP-MRF) // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2008.

134. Werner T. How to Compute Primal Solution from Dual One in MAP Inference in MRF? // Intl. Jr. on Control Systems and Computers. 2011. no. 2.

135. Woodford O. J., Rother C., Kolmogorov V. A Global Perspective on MAP Inference for Low-Level Vision // International Conference on Computer Vision (ICCV). 2009.

136. Yarkony J., Ihler A., Fowlkes C. Planar Cycle Covering Graphs // Uncertainty in Artificial Intelligence (UAI). 2011.

137. Yarkony J., Morshed R., Ihler A., Fowlkes C. Tightening MRF Relaxations with Planar Subproblems // Uncertainty in Artificial Intelligence (UAI). 2011.

138. Yedidia J. S., Freeman W. T., Weiss Y. Understanding belief propagation and its generalizations // Exploring Artificial Intelligence in the New Millennium / Ed. by G. Lakemeyer, B. Nebel. Morgan Kaufman, 2003.

139. Zach C., Haene C., Pollefeys M. What Is Optimized in Tight Convex Relaxations for Multi-Label Problems // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012.

A. Потоки и разрезы в сетях

Рассмотрим ориентированный граф $\bar{\mathcal{G}} = (\bar{\mathcal{V}}, \bar{\mathcal{E}})$, где $\bar{\mathcal{V}}$ – множество вершин, $\bar{\mathcal{E}}$ – множество ребер. Каждому ориентированному ребру (дуге) $(i, j) \in \bar{\mathcal{E}}$ соответствует неотрицательное число $c(i, j) \geqslant 0$ – *пропускная способность* (capacity). Пусть в графе есть две выделенные вершины: s – *исток*, t – *сток*. Граф $\bar{\mathcal{G}}$ вместе с введёнными пропускными способностями также называют *транспортной сетью*.

Потоком в сети назовем функцию $f : \bar{\mathcal{E}} \to \mathbb{R}$, такую что

1. $f(i, j) \leqslant c(i, j), \quad \forall (i \to j) \in \bar{\mathcal{E}}$;
2. $f(i, j) \geqslant 0, \quad \forall (i \to j) \in \bar{\mathcal{E}}$;
3. $\sum_{j:(i\to j)\in\bar{\mathcal{E}}} f(i, j) = \sum_{j:(j\to i)\in\bar{\mathcal{E}}} f(j, i), \quad \forall i \in \bar{\mathcal{V}} \setminus \{s, t\}$.

Величиной потока f назовем число

$$|f| = \sum_{i\in\bar{\mathcal{V}}} f(s, i) = \sum_{i\in\bar{\mathcal{V}}} f(i, t).$$

Задача о максимальном потоке в графе состоит в поиске потока f, обладающего максимальной величиной.

st-разрезом графа называется разбиение множества $\bar{\mathcal{V}}$ на два множества $\mathcal{S}$ и $\mathcal{T}$ ($\mathcal{S} \cup \mathcal{T} = \bar{\mathcal{V}}$, $\mathcal{S} \cap \mathcal{T} = \varnothing$), такое что $s \in \mathcal{S}$, $t \in \mathcal{T}$.

Величиной st-разреза называется сумма ёмкостей всех ребер, ведущих из $\mathcal{S}$ в $\mathcal{T}$:

$$\sum_{\substack{(i,j)\in\bar{\mathcal{E}} \\ j\in\mathcal{S},\, j\in\mathcal{T}}} c(i, j).$$

Задача о минимальном st-разрезе в графе состоит в поиске разреза $(\mathcal{S}, \mathcal{T})$, обладающего минимальной величиной.

Теорема Форда-Фалкерсона гласит, что величина максимального потока равна величине минимального st-разреза.

Известно, что задачи о максимальном потоке и о минимальном st-разрезе можно сформулировать как двойственные задачи линейного программирования (см. табл. A.1). В таких формулировках теорема Форда-Фалкерсона передоказывает сильную двойственность этих двух задач

Таблица А.1.: Задачи поиска максимального потока (слева) и минимального разреза (справа), сформулированные как задачи линейного программирования.

$$\begin{aligned}
\max_{f} \quad & \sum_{i:(s\to i)\in\bar{\mathcal{E}}} f(s,i) \\
\text{s.t.} \quad & f(i,j) \leqslant c(i,j), \quad \forall (i\to j)\in\bar{\mathcal{E}}; \\
& f(i,j) \geqslant 0, \quad \forall (i\to j)\in\bar{\mathcal{E}}; \\
& \sum_{j:(i\to j)\in\bar{\mathcal{E}}} f(i,j) = \sum_{j:(j\to i)\in\bar{\mathcal{E}}} f(j,i), \ \forall i\in\bar{\mathcal{V}}\setminus\{s,t\}.
\end{aligned}$$

$$\begin{aligned}
\min_{\boldsymbol{\mu},\boldsymbol{\nu}} \quad & \sum_{(i\to j)\in\bar{\mathcal{E}}} \mu_{ij} c(i,j) \\
\text{s.t.} \quad & \mu_{ij} \geqslant 0, \quad \forall (i\to j)\in\bar{\mathcal{E}}; \\
& \mu_{ij} \geqslant \nu_i - \nu_j, \quad \forall (i\to j)\in\bar{\mathcal{E}}; \\
& \nu_i \in \mathbb{R}, \ \forall i\in\bar{\mathcal{V}}\setminus\{s,t\}; \ \nu_s = 1, \ \nu_t = 0.
\end{aligned}$$

линейного программирования. Задача линейного программирования, соответствующая задаче поиска минимального разреза, специфична тем, что её решение совпадает с аналогичной задачей ЦЛП. Другими словами, верна теорема целочисленности (integrality theorem).

Отметим, что если нам известен максимальный поток, то найти минимальный разрез очень легко: например, можно к области истока $\mathcal{S}$ отнести все вершины, до которых существует путь по ненасыщенным ребрам сети ($f(i,j) < c(i,j)$). Построение максимального потока по минимальному разрезу – сложная задача.

Для решения задачи о максимальном потоке существует большое число алгоритмов. Классические алгоритмы Форда-Фалкерсона, проталкивание предпотока (push-relabel) описаны в книге “Алгоритмы: построение и анализ” (Т. Кормен и др.) [6]. Существует несколько специализированных алгоритмов, наиболее эффективных для задач, возникающих в компьютерном зрении: Бойкова-Колмогорова [23][1], IBFS [43][2].

Алгоритм IBFS в худшем случае выполняет $O(|\mathcal{V}|^2|\mathcal{E}|)$ операций (так же, как алгоритм проталкивания предпотока). На практике алгоритмы Бойкова-Колмогорова и IBFS при построении разрезов графов, возникающих в задачах компьютерного зрения, часто линейны по количеству вершин и рёбер графа.

[1] http://pub.ist.ac.at/~vnk/software/maxflow-v3.02.src.tar.gz
[2] http://www.cs.tau.ac.il/~sagihed/ibfs/