FMPE: Discriminatively Trained Features for Speech Recognition

Daniel Povey, Brian Kingsbury, Lidia Mangu, George Saon, Hagen Soltau, Geoffrey Zweig
IBM T.J. Watson Research Center, NY
ICASSP’2005

Presenter: Fang-Hui Chu
Outline (1/2)

- Introduction
- fMPE
  - High-dimensional feature generation
  - Acoustic context expansion
  - Feature projection
  - Training the matrix
  - Smoothing of update
- Calculating the Differential
  - Direct differential
  - Indirect differential
  - Model parameter differentials
  - Checks
Outline (2/2)

- Overview and General Considerations in fMPE Training
  - Overview
  - Dimension of high-dimensional features
  - Typical learning rates, and acoustic scaling
- CTS Experiments
- Call Center Experiments
- Conclusion
Introduction (1/2)

- fMPE applies the same objective function to the features

- The objective function of MPE is

\[
F_{\text{MPE}}(\lambda) = \sum_{r} \sum_{W_i \in W^r} p(W_i | O_r) A(W_i, W_r) \\
\approx \sum_{r} \sum_{W_i \in W^r} \frac{p(\lambda | W_i) P(W_i)}{\sum_{W_k \in W^r} p(\lambda | W_k) P(W_k)} A(W_i, W_r)
\]
Introduction (2/2)

MPE: Final Auxiliary Function

\[ H_{MPE} (\lambda, \bar{\lambda}) = \sum_r \sum_{q \in W_{lat}^r} \frac{\partial F_{MPE} (\lambda)}{\partial \log p(O_r | q)} _{\lambda = \bar{\lambda}} \log p(O_r | q) \]

\[ g_{MPE} (\lambda, \bar{\lambda}) = \sum_r \sum_{q \in W_{lat}^r} \sum_{t=e_q} \sum_m \gamma_{qm}^{r,MPE} \gamma_q^r (t) \log N(o_r(t), \mu_m, \Sigma_m) \]
fMPE (1/3)

- **Acoustic context expansion**
  
  \[ y_t = o_t + M h_t \]
  
  - \( o_t \) is the old feature
  - \( h_t \) is the high-dimensional feature
  - \( M \) is the transform matrix initialized to zero (starting point)

- **Feature projection**

  \[ d = 700,000 \]
fMPE (2/3)

High-dimension feature generation

- Transform the features into a very high dimensional space
- A set of Gaussians is created by likelihood-based clustering of the Gaussians in the acoustic model an appropriate size (100,000)
- On each frame, the Gaussian likelihoods are evaluated with no priors, and a vector of posteriors is formed
- Very quickly (less than 0.1xRT)
  - Only evaluate the 100 most likely clusters
  - Tricky→ Cluster the Gaussians to 2000 cluster centers
fMPE (3/3)

- Acoustic context expansion
  - The vector is further expanded with left and right acoustic context

- Feature projection
  - The high dimensional features are projected down to the dimension of the original features $x_t$ and added to them, so

  $$y_t = x_t + Mh_t$$

  Initializing $M$ to zero gives a reasonable starting point for training, i.e. the original features.
fMPE – auxiliary function

\[
g_{MPE}(M, \overline{M}) = \sum_{q \in W^r_{lattice}} \sum_{t=s_q}^{e_q} \sum_m \gamma_q \gamma_{qm}(t) \log N(y_t, \mu_m, \Sigma_m) \quad \left[ y_t = o_t + Mh_t \right]
\]

\[
= \sum_{q \in W^r_{lattice}} \sum_{t=s_q}^{e_q} \sum_m \gamma_q \gamma_{qm}(t) \log N(o_t + Mh_t, \mu_m, \Sigma_m)
\]

\[
\approx -\frac{1}{2} \sum_{q \in W^r_{lattice}} \sum_{t=s_q}^{e_q} \sum_m \gamma_q \gamma_{qm}(t)(o_t + Mh_t - \mu_m)^T \Sigma_m^{-1}(o_t + Mh_t - \mu_m)
\]

\[
\approx -\frac{1}{2} \sum_{q \in W^r_{lattice}} \sum_{t=s_q}^{e_q} \sum_m \gamma_q \gamma_{qm}(t) \sum_{i=1}^{D} \frac{(o_{t,i} + M_i h_t - \mu_{m,i})^2}{\sigma_{m,i}^2}
\]

\[
\therefore g_{MPE}(M_i, \overline{M}) \approx -\frac{1}{2} \sum_{q \in W^r_{lattice}} \sum_{t=s_q}^{e_q} \sum_m \gamma_q \gamma_{qm}(t) \frac{(o_{t,i} + M_i h_t - \mu_{m,i})^2}{\sigma_{m,i}^2}
\]

\[
\approx -\frac{1}{2} \sum_{q \in W^r_{lattice}} \sum_{t=s_q}^{e_q} \sum_m \gamma_q \gamma_{qm}(t) \frac{2(o_{t,i} - \mu_{m,i}) M_i h_t + (M_i h_t)^2}{\sigma_{m,i}^2}
\]
fMPE – differentiated w.r.t each row

\[
\frac{\partial g_{MPE}(M_i, \overline{M})}{\partial M_i} = -\frac{1}{2} \sum_{q \in W_{lattice}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{MPE} \gamma_{qm}(t) \left( \frac{2(o_{t,i} - \mu_{m,i})h_t}{\sigma_{m,i}^2} + \frac{2(M_i h_t)h_t}{\sigma_{m,i}^2} \right)
\]

\[
\therefore \frac{\partial g_{MPE}(M_i, \overline{M})}{\partial M_i} = 0 \Rightarrow \sum_{q \in W_{lattice}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{MPE} \gamma_{qm}(t) \frac{(\mu_{m,i} - o_{t,i})h_t^T}{\sigma_{m,i}^2}
\]

\[
= M_i \sum_{q \in W_{lattice}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{MPE} \gamma_{qm}(t) h_t h_t^T
\]

in such high dimensions accumulating squared statistics would be impractical.
fMPE – differentiated w.r.t each element

\[ g_{\text{MPE}} (M_i, \overline{M}) \approx -\frac{1}{2} \sum_{q \in \text{W}_{\text{lattice}}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{\text{MPE}} \gamma_{qm} (t) \frac{(o_{t,i} + M_i h_t - \mu_{m,i})^2}{\sigma_{m,i}^2} \]

\[ \approx -\frac{1}{2} \sum_{q \in \text{W}_{\text{lattice}}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{\text{MPE}} \gamma_{qm} (t) \frac{2(o_{t,i} - \mu_{m,i})M_i h_t + (M_i h_t)^2}{\sigma_{m,i}^2} \]

\[ g_{\text{MPE}} (M_{i,j}, \overline{M}) \approx -\frac{1}{2} \sum_{q \in \text{W}_{\text{lattice}}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{\text{MPE}} \gamma_{qm} (t) \frac{2(o_{t,i} - \mu_{m,i})M_{i,j} h_{t,j} + (M_i h_t)^2}{\sigma_{m,i}^2} \]

\[ \frac{\partial g_{\text{MPE}} (M_{i,j}, \overline{M})}{\partial M_{i,j}} = -\frac{1}{2} \sum_{q \in \text{W}_{\text{lattice}}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{\text{MPE}} \gamma_{qm} (t) \frac{2(o_{t,i} - \mu_{m,i})h_{t,j} + 2(M_i h_t)h_{t,j}}{\sigma_{m,i}^2} \]

\[ \therefore \frac{\partial g_{\text{MPE}} (M_{i,j}, \overline{M})}{\partial M_{i,j}} = 0 \Rightarrow \sum_{q \in \text{W}_{\text{lattice}}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{\text{MPE}} \gamma_{qm} (t) \frac{(\mu_{m,i} - o_{t,i})h_{t,j}}{\sigma_{m,i}^2} \]

\[ = M_i \sum_{q \in \text{W}_{\text{lattice}}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_q^{\text{MPE}} \gamma_{qm} (t) \frac{h_{t,j} h_t}{\sigma_{m,i}^2} \]

Need to solve the linear equations
fMPE – differentiated w.r.t each element (on paper)

\[ g_{MPE}(M, \bar{M}) = -\frac{1}{2} \sum_{q \in W^r_{lattice}} \sum_{t=e_q}^{t=s_q} \sum_{m} \gamma_q^{MPE} \gamma_{qm}(t) \sum_{i=1}^{D} \frac{(y_{t,i} - \mu_{m,i})^2}{\sigma_{m,i}^2} \]

where \( y_{t,i} = o_{t,i} + M_i h_t = o_{t,i} + \sum_j M_{i,j} h_{t,j} \)

\[ \frac{\partial g_{MPE}(M, \bar{M})}{\partial M_{i,j}} = \sum_{y_{t,i}} \frac{\partial g_{MPE}(y_{t,i}, \bar{M})}{\partial y_{t,i}} \frac{\partial y_{t,i}}{\partial M_{i,j}} = \sum_{y_{t,i}} \frac{\partial g_{MPE}(y_{t,i}, \bar{M})}{\partial y_{t,i}} h_{t,j} \]

\[ g_{MPE}(y_{t,i}, \bar{M}) = -\frac{1}{2} \sum_{q \in W^r_{lattice}} \sum_{m} \gamma_q^{MPE} \gamma_{qm}(t) \frac{y_{t,i}^2 - 2\mu_{m,i}y_{t,i}}{\sigma_{m,i}^2} \]

\[ \frac{\partial g_{MPE}(y_{t,i}, \bar{M})}{\partial y_{t,i}} = \sum_{q \in W^r_{lattice}} \sum_{m} \gamma_q^{MPE} \gamma_{qm}(t) \frac{\mu_{m,i} - y_{t,i}}{\sigma_{m,i}^2} \]

\[ \therefore \frac{\partial g_{MPE}(M, \bar{M})}{\partial M_{i,j}} = \sum_{q \in W^r_{lattice}} \sum_{t=e_q}^{t=s_q} \sum_{m} \gamma_q^{MPE} \gamma_{qm}(t) \frac{\mu_{m,i} - y_{t,i}}{\sigma_{m,i}^2} h_{t,j} \]
fMPE (1/5)

Training the matrix

The matrix is trained by linear methods (gradient descent)

\[
M_{i,j} := M_{i,j} + \nu_{i,j} \frac{\partial g_{MPE}(M, \bar{M})}{\partial M_{i,j}} \\
:= M_{i,j} + \nu_{i,j} \left( \sum_{q \in W_{\text{lattice}}} \sum_{t=s_q}^{t=e_q} \sum_{m} \gamma_{q}^{MPE} \gamma_{q,m}(t) \frac{\mu_{m,i} - y_{t,i}}{\sigma_{m,i}^2} h_{t,j} \right)
\]

The transformed features have the maximal average accuracy on current models. But, how about on the models re-trained by these transformed features?
We should modify the auxiliary function to

\[
g_{\text{MPE}}(M, \overline{M}) = \sum_{q \in W_{\text{lattice}}} \sum_{t} \sum_{m} \gamma_{q}^{\text{MPE}} \gamma_{qm}(t) \log N(y_{t}, \tilde{\mu}_{m}, \tilde{\Sigma}_{m})
\]

where \( \tilde{\mu}_{m} = \frac{\sum_{t} \tilde{y}_{m}(t) y_{t}}{\sum_{t} \tilde{y}_{m}(t)} \) and \( \tilde{\Sigma}_{m} = \frac{\sum_{t} \tilde{y}_{m}(t) (y_{t} - \tilde{\mu}_{m})(y_{t} - \tilde{\mu}_{m})^{T}}{\sum_{t} \tilde{y}_{m}(t)} \)

is the newly model parameters will be trained using transformed features \( y_{t} \)

\[
\therefore g_{\text{MPE}}(M_{i}, \overline{M}_{i}) = -\frac{1}{2} \sum_{t} \sum_{q \in W_{\text{lattice}}} \sum_{m} \gamma_{q}^{\text{MPE}} \gamma_{qm}(t) \left( \log \sigma_{m,i}^{2} + \frac{(y_{t,i} - \tilde{\mu}_{m,i})^{2}}{\sigma_{m,i}^{2}} \right)
\]

where \( \tilde{\mu}_{m,i} = \frac{\sum_{t} \tilde{y}_{m}(t) y_{t,i}}{\sum_{t} \tilde{y}_{m}(t)} \), and \( \tilde{\sigma}_{m,i}^{2} = \frac{\sum_{t} \tilde{y}_{m}(t)(y_{t,i} - \tilde{\mu}_{m,i})^{2}}{\sum_{t} \tilde{y}_{m}(t)} \)

\[
\frac{\partial g_{\text{MPE}}(M, \overline{M})}{\partial y_{t,i}} = \sum_{q \in W_{\text{lattice}}} \sum_{m} \gamma_{q}^{\text{MPE}} \gamma_{qm}(t) \frac{\tilde{\mu}_{m,i} - y_{t,i}}{\tilde{\sigma}_{m,i}^{2}} + \sum_{m} \frac{\partial g_{\text{MPE}}(y_{t,j}, M)}{\partial \tilde{\mu}_{m,i}} \frac{\partial \tilde{\mu}_{m,i}}{\partial y_{t,i}} + \sum_{m} \frac{\partial g_{\text{MPE}}(y_{t,j}, M)}{\partial \tilde{\sigma}_{m,i}^{2}} \frac{\partial \tilde{\sigma}_{m,i}^{2}}{\partial y_{t,i}}
\]

direct

indirect
fMPE (3/5)

\[
\begin{align*}
\therefore \frac{\partial \tilde{\mu}_{m,i}}{\partial y_{t,i}} &= \sum_{t^*} \tilde{\gamma}_m(t^*) \text{ and } \frac{\partial \tilde{\sigma}^2_{m,i}}{\partial y_{t,i}} = 2 \sum_{t^*} \gamma_m(t^*) (y_{t,i} - \tilde{\mu}_{m,i}) \\
\therefore \sum_{m} \frac{\partial g_{\text{MPE}}(M, \bar{M})}{\partial \tilde{\mu}_{m,i}} \frac{\partial \tilde{\mu}_{m,i}}{\partial y_{t,i}} &= \sum_{m} \frac{\tilde{\gamma}_m(t)}{\gamma_m(t^*)} \frac{\partial g_{\text{MPE}}(M, \bar{M})}{\partial \tilde{\mu}_{m,i}} \\
\sum_{m} \frac{\partial g_{\text{MPE}}(M, \bar{M})}{\partial \tilde{\sigma}^2_{m,i}} \frac{\partial \tilde{\sigma}^2_{m,i}}{\partial y_{t,i}} &= \sum_{m} 2 \frac{\partial g_{\text{MPE}}(M, \bar{M})}{\partial \tilde{\sigma}^2_{m,i}} \sum_{t^*} \frac{\tilde{\gamma}_m(t)}{\gamma_m(t^*)} (y_{t,i} - \tilde{\mu}_{m,i}) \\
\frac{\partial g_{\text{MPE}}(M, \bar{M})}{\partial \tilde{\mu}_{m,i}} &= \frac{1}{\tilde{\sigma}^2_{m,i}} \sum_{t'} \sum_{q \in W_{\text{lattice}}^r} \gamma_{q, \text{MPE}}(t') (y_{t',i} - \tilde{\mu}_{m,i}) \\
\frac{\partial g_{\text{MPE}}(M, \bar{M})}{\partial \tilde{\sigma}^2_{m,i}} &= \frac{1}{2} \sum_{t'} \sum_{q \in W_{\text{lattice}}^r} \gamma_{q, \text{MPE}}(t') \left((y_{t',i} - \tilde{\mu}_{m,i})^2 \tilde{\sigma}^{-4}_{m,i} - \tilde{\sigma}^{-2}_{m,i}\right)
\end{align*}
\]
fMPE (4/5)

Training the matrix

The parameter-specific learning rates are:

\[ v_{ij} = \frac{\sigma_i}{E(p_{ij} + n_{ij})} \]

Since \( \frac{\partial F}{\partial M_{ij}} = p_{ij} - n_{ij} \), the most each \( M_{ij} \) can change is \( 1/E \) standard deviations

The most any given feature element \( y_{ti} \) can change is \( n/E \) standard deviations, where \( n \) is the number of acoustic contexts by which the vector \( H_t \) has been expanded (e.g. \( n=7 \))

\[ \frac{\partial F}{\partial M_{ij}} = \sum_{t=1}^{T} \frac{\partial F}{\partial y_{ti}} h_{ij} \quad p_{ij} = \sum_{t=1}^{T} \max(\frac{\partial F}{\partial y_{ti}} h_{ij}, 0) \quad n_{ij} = \sum_{t=1}^{T} \max(-\frac{\partial F}{\partial y_{ti}} h_{ij}, 0) \]
Smoothing of update

To prevent over-training of parameters that cannot be estimated robustly, a modification is made

\[
n_{ij} := n_{ij} + 0.5\tau(p_{ij} + n_{ij})/c_{ij}
\]

\[
p_{ij} := p_{ij} + 0.5\tau(p_{ij} + n_{ij})/c_{ij}
\]

\[
c_{ij} = \sum_{t=1}^{T} h_{ij}
\]

which is similar to the number of nonzero points available in estimating the differential \( \frac{\partial F}{\partial M_{ij}} \)

\[
c_{ij} = (\sum_{t=1}^{T}|d_{ij}(t)|)^2 / \sum_{t=1}^{T} d_{ij}(t) \]

\[
d_{ij}(t) = \frac{\partial F}{\partial y_{ti}} h_{ij}
\]

which is the number of points that would have the same expected ratio of squared sum of absolute values to sum-of-squares if it were Gaussian distributed with zero mean

Smoothing may slightly improve results, on the order of 0.1% absolute; generally this is done with \( \tau \approx 100 \)
A useful check that no implementation errors have been made is

adding a small quantity to all the features in some dimension should not affect the MPE objective function, as long as it is done in both training and test.

\[
\sum_{t=1}^{T} \frac{\partial F}{\partial y_{ti}}^{\text{direct}} + \frac{\partial F}{\partial y_{ti}}^{\text{indirect}} = 0
\]

\[
\sum_{t=1}^{T} y_{ti} \frac{\partial F}{\partial y_{ti}}^{\text{direct}} + y_{ti} \frac{\partial F}{\partial y_{ti}}^{\text{indirect}} = 0
\]
\[
\sum_t \frac{\partial g_{MPE}(M, M)}{\partial y_{t,i}} = \\
\sum_t \sum_{q \in \mathcal{W}_m} \sum_{m} \gamma_q^{MPE} \gamma_q^{m} (t) \left( \frac{\bar{\mu}_{m,i} - y_{t,i}}{\bar{\sigma}^2_{m,i}} \right) + \sum_t \sum_{m} \gamma_m(t) \left( \frac{1}{\bar{\sigma}^2_{m,i}} \left( \sum_{q \in \mathcal{W}_m} \gamma_q^{MPE} \gamma_q^{m}(t') \left( y_{t',i} - \bar{\mu}_{m,i} \right) \right) \right) \\
+ \sum_t \sum_{m} \sum_{t'} \left( \sum_{q \in \mathcal{W}_m} \gamma_q^{MPE} \gamma_q^{m}(t'') \left( \frac{\left( y_{t'',i} - \bar{\mu}_{m,i} \right)^2}{\bar{\sigma}^2_{m,i}} - 1 \right) \right) \sum_{t'} \gamma_m(t') \left( y_{t',i} - \bar{\mu}_{m,i} \right) \\
= \sum_m \left( \frac{1}{\bar{\sigma}^2_{m,i}} \sum_{t'} \sum_{q \in \mathcal{W}_m} \gamma_q^{MPE} \gamma_q^{m}(t'') \left( \frac{\left( y_{t'',i} - \bar{\mu}_{m,i} \right)^2}{\bar{\sigma}^2_{m,i}} - 1 \right) \right) \sum_t \gamma_m(t') y_{t,i} - \sum_t \gamma_m(t') \bar{\mu}_{m,i} \\
= 0 \left( \sum_t \gamma_m(t') y_{t,i} \approx \bar{\mu}_{m,i} \right)
\]
Overview and general considerations in fMPE training

Overview

Each iteration of fMPE training involves three passes over the date:

- one to accumulate normal MPE statistics
- second to accumulate fMPE statistics (chiefly the quantities $n_{ij}$ and $p_{ij}$)
- third to do an ML update with the newly transformed data

Dimension of high-dimensional features

Experiments on call center data suggest that it is probably good to use as high a dimension as possible until there is insufficient data for each parameter and data-learning becomes an issue
Overview and general considerations in fMPE training

Typical learning rates, and acoustic scaling

The values of $E$ used in the CTS experiments reported here are 0.96 for the speaker independent system

- 5 acoustic contexts in the high dimensional features
- 1.44 for the speaker adapted system
- 7 acoustic contexts in the high dimensional features

The call-center experiments also use 7 contexts and $E = 1.44$

To prevent the fMPE transform from attempting to generally strengthen or weaken the acoustic model relative to the LM

- the differential of the MPE criterion w.r.t a scaling of all the acoustic likelihoods was calculated and the LM weight was tuned until this was close to zero
Conversational telephone speech (CTS) experiments

- A set of statistics (corresponding to the denominator statistics in MMI training) is also accumulated so that I-smoothing can back off to an MMI rather than an ML estimate.

- Training is on 2300h of telephone speech data.

- Both systems used cross-word phonetic context, and PLP features with LDA+MLLT projections to 40 dimensions (SI) and 39 (adapted).
Conversational telephone speech (CTS) experiments

- The SI system is a quinphone system with 8k states and 150k Gaussians
- The high-dimensional features are posteriors of 64k clustered Gaussians with five contexts
- The transform is trained with 1/5 of the training data
- fMPE+MPE is better by 1.0% than MPE alone
Conversational telephone speech (CTS) experiments

- The adapted system has 7-phone context, 22k states and 850k Gaussians, training and testing on VTLN+fMLLR features.

- The $h_t$ are posteriors of 100k Gaussians, with seven contexts (700k dimensions total).

- The transform is trained on all the data.

- In this case fMPE alone is better than MPE alone.
Call center experiments

- Training is on 300h of speech
- The models have 11-phone left phonetic context, 4k states and 97k Gaussians
- Test data is 6 hours long
- Features are PLP projected with LDA+MLLT to 40 dimensions
- High dimensional features are 32k Gaussian posteriors with 7 contexts (224,000 dimensions)
- MPE is with backoff to MMI as above
- The fMPE+MPE results on call-center data are an impressive 5.1% better than the ML baseline and 1.7% better than MPE alone.
Conclusion

fMPE is a novel and effective way to apply discriminative training to features rather than models.

This makes possible things that are not possible with normal discriminative training, such as building a system on the new features and iterating the process.
fMPE-check1

```
check_14_dir[0]=1.0371757236e+004  
check_14_indir_m[0]=-1.0371757236e+004  
check_14_indir_v[0]=-2.9030436995e+002  
check_14_dir[1]=-2.7145537911e+003  
check_14_indir_m[1]=-2.7145537911e+003  
check_14_indir_v[1]=3.0971625056e+001  
check_14_dir[2]=-4.6783840780e+002  
check_14_indir_m[2]=4.6783840780e+002  
check_14_indir_v[3]=-1.30431079e+002  
check_14_dir[4]=1.414851846e+003  
check_14_indir_m[4]=-1.414851846e+003  
check_14_indir_v[4]=5.5481249975e+000  
check_14_indir_m[5]=7.977883357e+002  
check_14_indir_v[5]=3.899625056e+001  
check_14_dir[6]=3.3513488117e+002  
check_14_indir_m[6]=-2.3513488117e+002  
check_14_indir_v[6]=-1.865977121e+002  
check_14_dir[7]=5.695810251e+002  
check_14_indir_m[7]=-5.695810251e+002  
check_14_indir_v[7]=-4.9958412737e+001  
check_14_dir[8]=3.4810229854e+002  
check_14_indir_m[8]=3.4810229854e+002  
check_14_indir_v[8]=-7.3710301259e+001  
check_14_dir[9]=-1.9179030716e+003  
check_14_indir_m[9]=-1.9179030716e+003  
check_14_indir_v[9]=-4.4725724363e+001  
check_14_dir[10]=7.4367252790e+002  
check_14_indir_m[10]=7.4367252790e+002  
check_14_indir_v[10]=-1.5727315623e+001  
check_14_indir_v[11]=-1.118616687e+001  
check_14_dir[12]=-1.0044030731e+004  
check_14_indir_m[12]=1.0044030731e+004  
check_14_indir_v[12]=-2.8314310300e+001  
```

Cancel out!
\[
\sum_t y_{t,i} \frac{\partial g_{\text{MPE}}(M, \bar{M})}{\partial y_{t,i}} =
\sum_t y_{t,i} \sum_{q \in W_{\text{lattice}}} \sum_m \gamma_q^\text{MPE} \gamma_{qm}(t) \frac{\bar{\mu}_{m,i} - y_{t,i}}{\bar{\sigma}_{m,i}^2} + \sum_t y_{t,i} \sum_m \frac{\bar{\gamma}_m(t)}{\bar{\sigma}_{m,i}^2} \left( \sum_{t'} \sum_{q \in W_{\text{lattice}}} \gamma_q^\text{MPE} \gamma_{qm}(t') (y_{t',i} - \bar{\mu}_{m,i}) \right) \\
+ \sum_t y_{t,i} \sum_m \left( \sum_{t'} \sum_{q \in W_{\text{lattice}}} \gamma_q^\text{MPE} \gamma_{qm}(t') \left( (y_{t',i} - \bar{\mu}_{m,i})^2 \bar{\sigma}_{m,i}^{-4} - \bar{\sigma}_{m,i}^{-2} \right) \frac{\bar{\gamma}_m(t)}{\bar{\gamma}_m(t')} \right) (y_{t,i} - \bar{\mu}_{m,i}) \\
= \sum_m \left( \frac{1}{\bar{\sigma}_{m,i}^2} \sum_{t'} \sum_{q \in W_{\text{lattice}}} \gamma_q^\text{MPE} \gamma_{qm}(t') \left( \frac{(y_{t',i} - \bar{\mu}_{m,i})^2}{\bar{\sigma}_{m,i}^2} - 1 \right) \right) \left( \sum_t \frac{\bar{\gamma}_m(t)}{\bar{\gamma}_m(t')} y_{t,i} y_{t,i} - \sum_t \frac{\bar{\gamma}_m(t)}{\bar{\gamma}_m(t')} y_{t,i} \bar{\mu}_{m,i} \right) \\
\therefore \sum_t \sum_{t'} \frac{\bar{\gamma}_m(t)}{\bar{\gamma}_m(t')} y_{t,i} \bar{\mu}_{m,i} = \bar{\mu}_{m,i}^2 \quad \text{and} \quad \sum_t \sum_{t'} \frac{\bar{\gamma}_m(t)}{\bar{\gamma}_m(t')} y_{t,i} y_{t,i} = \bar{\sigma}_{m,i}^2 + \bar{\mu}_{m,i}^2 \\
= \sum_m \left( \sum_{t'} \sum_{q \in W_{\text{lattice}}} \gamma_q^\text{MPE} \gamma_{qm}(t') \left( \frac{(y_{t',i} - \bar{\mu}_{m,i})^2}{\bar{\sigma}_{m,i}^2} - 1 \right) \right) 
\]
fMPE – check2

\[
\sum_{t'} \sum_{m} \sum_{q \in W_{\text{lattice}}} \gamma^\text{MPE}_{q} \gamma_{qm}(t'') \left( \frac{(y_{t',i} - \tilde{\mu}_{m,i})^2}{\tilde{\sigma}^2_{m,i}} - 1 \right)
\]

\[
= \sum_{t'} \sum_{m} \sum_{q \in W_{\text{lattice}}} \gamma^\text{MPE}_{q} \gamma_{qm}(t'') \left( \frac{(y_{t',i} - \tilde{\mu}_{m,i})^2}{\tilde{\sigma}^2_{m,i}} - \sum_{t''} \sum_{m} \sum_{q \in W_{\text{lattice}}} \gamma^\text{MPE}_{q} \gamma_{qm}(t'') \right)
\]

\[
= \sum_{m} \sum_{t'} \sum_{q \in W_{\text{lattice}}} \gamma^\text{MPE}_{q} \gamma_{qm}(t'') \sum_{t''} \sum_{m} \sum_{q \in W_{\text{lattice}}} \gamma^\text{MPE}_{q} \gamma_{qm}(t'') \left( \frac{(y_{t',i} - \tilde{\mu}_{m,i})^2}{\tilde{\sigma}^2_{m,i}} \right)
\]

\[
= \sum_{m} \frac{\sigma^2_{m,i}}{\tilde{\sigma}^2_{m,i}} \sum_{t'} \sum_{q \in W_{\text{lattice}}} \gamma^\text{MPE}_{q} \gamma_{qm}(t'') \rightarrow 0 \quad \left( \frac{\sigma^2_{m,i}}{\tilde{\sigma}^2_{m,i}} \rightarrow 1 \right)
\]